(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 February 2009 (05.02.2009)

PCT

(10) International Publication Number
WO 2009/016631 A2

(51) International Patent Classification:
G06F 17/22 (2006.01)

(21) International Application Number:
PCT/IL2008/001051

(22) International Filing Date: 31 July 2008 (31.07.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/953,209 1 August 2007 (01.08.2007) US

(71) Applicant *(for all designated States except US)*: **GINGER SOFTWARE, INC.** [US/US]; 405 Waltham Street, #371, Lexington, Massachusetts 02421 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KAROV ZANGVIL, Yael** [IL/IL]; 15 Kfar Saba Street, Neve Zedek, 65147 Tel-aviv (IL). **ZANGVIL, Avner** [IL/IL]; 15 Kfar Saba Street, Neve Zedek, 65147 Tel-aviv (IL).

(74) Agents: **SANFORD T. COLB & CO.** et al.; P.O. Box 2273, 76122 Rehovot (IL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

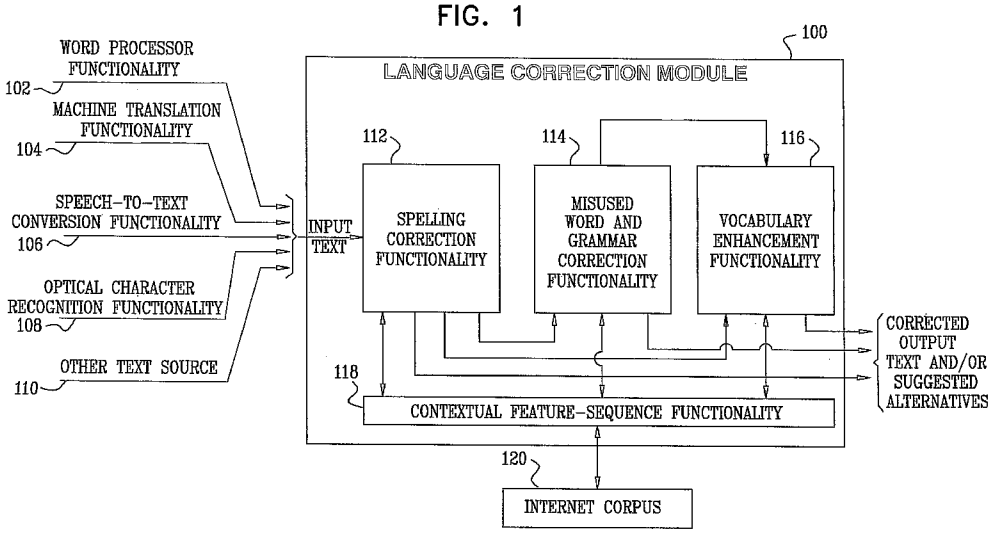
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

(54) Title: AUTOMATIC CONTEXT SENSITIVE LANGUAGE CORRECTION AND ENHANCEMENT USING AN INTERNET CORPUS

FIG. 1



(57) Abstract: A computer-assisted language correction system including spelling correction functionality, misused word correction functionality, grammar correction functionality and vocabulary enhancement functionality utilizing contextual feature-sequence functionality employing an internet corpus.

WO 2009/016631 A2

AUTOMATIC CONTEXT SENSITIVE LANGUAGE CORRECTION AND ENHANCEMENT USING AN INTERNET CORPUS

REFERENCE TO RELATED APPLICATIONS

Reference is hereby made to U.S. Provisional Patent Application Serial No. 60/953,209, filed August 1, 2007, entitled METHODS FOR CONTEXT SENSITIVE ERROR DETECTION AND CORRECTION, the disclosure of which is hereby incorporated by reference and priority of which is hereby claimed pursuant to 37 CFR 1.78(a) (4) and (5)(i).

FIELD OF THE INVENTION

The present invention relates to computer-assisted language correction generally.

BACKGROUND OF THE INVENTION

The following publications are believed to represent the current state of the art:

U.S. Patent Nos. 5,659,771; 5,907,839; 6,424,983; 7,296,019; 5,956,739 and 4,674,065

U.S. Published Patent Application Nos. 2006/0247914 and 2007/0106937;

SUMMARY OF THE INVENTION

The present invention seeks to provide improved systems and functionalities for computer-assisted language correction.

There is thus provided in accordance with a preferred embodiment of the present invention a computer-assisted language correction system including an alternatives generator, generating on the basis of an input sentence a text-based representation providing multiple alternatives for each of a plurality of words in the sentence, a selector for selecting among at least the multiple alternatives for each of the plurality of words in the sentence, based at least partly on an internet corpus, and a correction generator operative to provide a correction output based on selections made by the selector.

Preferably, the selector is operative to make the selections based on at least one of the following correction functions: spelling correction, misused word correction, grammar correction and vocabulary enhancement.

In accordance with a preferred embodiment of the present invention the selector is operative to make the selections based on at least two of the following correction functions: spelling correction, misused word correction, grammar correction; and vocabulary enhancement. Additionally, the selector is operative to make the selections based on at least one of the following time ordering of corrections: spelling correction prior to at least one of misused word correction, grammar correction and vocabulary enhancement, and misused word correction and grammar correction prior to vocabulary enhancement.

Additionally or alternatively, the input sentence is provided by one of the following functionalities: word processor functionality, machine translation functionality, speech-to-text conversion functionality, optical character recognition functionality and instant messaging functionality, and the selector is operative to make the selections based on at least one of the following correction functions: misused word correction, grammar correction and vocabulary enhancement.

Preferably, the correction generator includes a corrected language input generator operative to provide a corrected language output based on selections made by the selector without requiring user intervention. Additionally or alternatively, the grammar correction functionality includes at least one of punctuation, verb inflection, single/plural, article and preposition correction functionalities.

In accordance with a preferred embodiment of the present invention the grammar correction functionality includes at least one of replacement, insertion and omission correction functionalities.

Preferably, the selector includes context based scoring functionality operative to rank the multiple alternatives, based at least partially on contextual feature-sequence (CFS) frequencies of occurrences in an internet corpus. Additionally, the context based scoring functionality is also operative to rank the multiple alternatives based at least partially on normalized CFS frequencies of occurrences in the internet corpus.

There is also provided in accordance with another preferred embodiment of the present invention a computer-assisted language correction system including at least one of spelling correction functionality, misused word correction functionality, grammar correction functionality and vocabulary enhancement functionality, and contextual feature-sequence functionality cooperating with at least one of the spelling correction functionality; the misused word correction functionality, grammar correction functionality and the vocabulary enhancement functionality and employing an internet corpus.

Preferably, the grammar correction functionality includes at least one of punctuation, verb inflection, single/plural, article and preposition correction functionalities. Additionally or alternatively, the grammar correction functionality includes at least one of replacement, insertion and omission correction functionalities.

In accordance with a preferred embodiment of the present invention the computer-assisted language correction system includes at least two of the spelling correction functionality, the misused word correction functionality, the grammar correction functionality and the vocabulary enhancement functionality, and the contextual feature-sequence functionality cooperates with at least two of the spelling

correction functionality, the misused word correction functionality, the grammar correction functionality and the vocabulary enhancement functionality, and employs an internet corpus.

Preferably, the computer-assisted language correction system also includes at least three of the spelling correction functionality, the misused word correction functionality; the grammar correction functionality and the vocabulary enhancement functionality and the contextual feature-sequence functionality cooperates with at least three of the spelling correction functionality, the misused word correction functionality, the grammar correction functionality and the vocabulary enhancement functionality, and employs an internet corpus.

In accordance with a preferred embodiment of the present invention the computer-assisted language correction system also includes the spelling correction functionality, the misused word correction functionality, the grammar correction functionality and the vocabulary enhancement functionality, and the contextual feature-sequence functionality cooperates with the spelling correction functionality, the misused word correction functionality, the grammar correction functionality and the vocabulary enhancement functionality, and employs an internet corpus.

Preferably, the correction generator includes a corrected language generator operative to provide a corrected language output based on selections made by the selector without requiring user intervention.

There is further provided in accordance with yet another preferred embodiment of the present invention a computer-assisted language correction system including an alternatives generator, generating on the basis of a language input a text-based representation providing multiple alternatives for each of a plurality of words in the sentence, a selector for selecting among at least the multiple alternatives for each of the plurality of words in the language input, based at least partly on a relationship between selected ones of the multiple alternatives for at least some of the plurality of words in the language input and a correction generator operative to provide a correction output based on selections made by the selector.

Preferably, the language input includes at least one of an input sentence and an input text. Additionally or alternatively, the language input is speech and the

generator converts the language input in speech to a text-based representation providing multiple alternatives for a plurality of words in the language input.

In accordance with a preferred embodiment of the present invention the language input is at least one of a text input, an output of optical character recognition functionality, an output of machine translation functionality and an output of word processing functionality, and the generator converts the language input in text to a text-based representation providing multiple alternatives for a plurality of words in the language input.

Preferably, the selector is operative to make the selections based on at least two of the following correction functions: spelling correction, misused word correction, grammar correction and vocabulary enhancement. Additionally, the selector is operative to make the selections based on at least one of the following time ordering of corrections: spelling correction prior to at least one of misused word correction, grammar correction and vocabulary enhancement, and misused word correction and grammar correction prior to vocabulary enhancement.

In accordance with a preferred embodiment of the present invention the language input is speech and the selector is operative to make the selections based on at least one of the following correction functions: misused word correction, grammar correction and vocabulary enhancement.

Preferably, the selector is operative to make the selections by carrying out at least two of the following functions: selection of a first set of words or combinations of words which include less than all of the plurality of words in the language input for an initial selection, thereafter ordering elements of the first set of words or combinations of words to establish priority of selection and thereafter when selecting among the multiple alternatives for an element of the first set of words, choosing other words, but not all, of the plurality of words as a context to influence the selecting. Additionally or alternatively, the selector is operative to make the selections by carrying out the following function: when selecting for an element having at least two words, evaluating each of the multiple alternatives for each of the at least two words in combination with each of the multiple alternatives for each other of the at least two words.

In accordance with a preferred embodiment of the present invention the correction generator includes a corrected language input generator operative to provide a corrected language output based on selections made by the selector without requiring user intervention.

There is even further provided in accordance with still another preferred embodiment of the present invention a computer-assisted language correction system including a misused-word suspector evaluating at least most of the words in an language input on the basis of their fit within a context of the language input and a correction generator operative to provide a correction output based at least partially on an evaluation performed by the suspector.

Preferably, the computer-assisted language correction system also includes an alternatives generator, generating on the basis of the language input, a text-based representation providing multiple alternatives for at least one of the at least most words in the language input and a selector for selecting among at least the multiple alternatives for each of the at least one of the at least most words in the language input, and the correction generator is operative to provide the correction output based on selections made by the selector. Additionally or alternatively, the computer-assisted language correction system also includes a suspect word output indicator indicating an extent to which at least some of the at least most of the words in the language input is suspect as a misused-word.

In accordance with a preferred embodiment of the present invention the correction generator includes an automatic corrected language generator operative to provide a corrected text output based at least partially on an evaluation performed by the suspector, without requiring user intervention.

Preferably, the language input is speech and the selector is operative to make the selections based on at least one of the following correction functions: misused word correction, grammar correction and vocabulary enhancement.

There is also provided in accordance with still another preferred embodiment of the present invention a computer-assisted language correction system including a misused-word suspector evaluating words in an language input, an alternatives generator, generating multiple alternatives for at least some of the words in

the language input evaluated as suspect words by the suspector, at least one of the multiple alternatives for a word in the language input being consistent with a contextual feature of the word in the language input in an internet corpus, a selector for selecting among at least the multiple alternatives and a correction generator operative to provide a correction output based at least partially on a selection made by the selector.

There is further provided in accordance with yet another preferred embodiment of the present invention a computer-assisted language correction system including a misused-word suspector evaluating words in an language input and identifying suspect words, an alternatives generator, generating multiple alternatives for the suspect words, a selector, grading each the suspect word as well as ones of the multiple alternatives therefor generated by the alternatives generator according to multiple selection criteria, and applying a bias in favor of the suspect word vis-à-vis ones of the multiple alternatives therefor generated by the alternatives generator and a correction generator operative to provide a correction output based at least partially on a selection made by the selector.

There is yet further provided in accordance with still another preferred embodiment of the present invention a computer-assisted language correction system including an alternatives generator, generating on the basis of an input multiple alternatives for at least one word in the input, a selector, grading each the at least one word as well as ones of the multiple alternatives therefor generated by the alternatives generator according to multiple selection criteria, and applying a bias in favor of the at least one word vis-à-vis ones of the multiple alternatives therefor generated by the alternatives generator, the bias being a function of an input uncertainty metric indicating uncertainty of a person providing the input, and a correction generator operative to provide a correction output based on a selection made by the selector.

There is even further provided in accordance with another preferred embodiment of the present invention a computer-assisted language correction system including an incorrect word suspector evaluating at least most of the words in a language input, the suspector being at least partially responsive to an input uncertainty metric indicating uncertainty of a person providing the input, the suspector providing a suspected incorrect word output, and an alternatives generator, generating a plurality of

alternatives for suspected incorrect words identified by the suspected incorrect word output, a selector for selecting among each suspected incorrect word and the plurality of alternatives generated by the alternatives generator, and a correction generator operative to provide a correction output based on a selection made by the selector.

There is also provided in accordance with yet another preferred embodiment of the present invention a computer-assisted language correction system including at least one of a spelling correction module, a misused-word correction module, a grammar correction module and a vocabulary enhancement module receiving a multi-word input and providing a correction output, each of the at least one of a spelling correction module, a misused-word correction module, a grammar correction module and a vocabulary enhancement module including an alternative word candidate generator including phonetic similarity functionality operative to propose alternative words based on phonetic similarity to a word in the input and to indicate a metric of phonetic similarity and character string similarity functionality operative to propose alternative words based on character string similarity to a word in the input and to indicate a metric of character string similarity for each alternative word, and a selector operative to select either a word in the output or an alternative word candidate proposed by the alternative word candidate generator by employing the phonetic similarity and character string similarity metrics together with context-based selection functionality.

There is even further provided in accordance with still another preferred embodiment of the present invention a computer-assisted language correction system including suspect word identification functionality, receiving a multi-word language input and providing a suspect word output which indicates suspect words, feature identification functionality operative to identify features including the suspect words, an alternative selector identifying alternatives to the suspect words, feature occurrence functionality employing a corpus and providing an occurrence output, ranking various features including the alternatives as to their frequency of use in the corpus, and a selector employing the occurrence output to provide a correction output, the feature identification functionality including feature filtration functionality including at least one of functionality for eliminating features containing suspected errors, functionality for negatively biasing features which contain words introduced in an earlier correction

iteration of the multi-word input and which have a confidence level below a confidence level predetermined threshold, and functionality for eliminating features which are contained in another feature having an frequency of occurrence above a predetermined frequency threshold.

Preferably, the selector is operative to make the selections based on at least two of the following correction functions: spelling correction, misused word correction, grammar correction and vocabulary enhancement. Additionally, the selector is operative to make the selections based on at least one of the following time ordering of corrections: spelling correction prior to at least one of misused word correction, grammar correction and vocabulary enhancement and misused word correction and grammar correction prior to vocabulary enhancement.

In accordance with a preferred embodiment of the present invention the language input is speech and the selector is operative to make the selections based on at least one of the following correction functions: grammar correction, and misused word correction and vocabulary enhancement.

Preferably, the correction generator includes a corrected language input generator operative to provide a corrected language output based on selections made by the selector without requiring user intervention.

In accordance with a preferred embodiment of the present invention the selector is also operative to make the selections based at least partly on a user input uncertainty metric. Additionally, the user input uncertainty metric is a function based on a measurement of the uncertainty of a person providing the input. Additionally or alternatively, the selector also employs user input history learning functionality.

There is still further provided in accordance with yet another preferred embodiment of the present invention a computer-assisted language correction system including suspect word identification functionality, receiving a multi-word language input and providing a suspect word output which indicates suspect words,feature identification functionality operative to identify features including the suspect words, an alternative selector identifying alternatives to the suspect words, occurrence functionality employing a corpus and providing an occurrence output, ranking features including the alternatives as to their frequency of use in the corpus, and a correction

output generator, employing the occurrence output to provide a correction output, the feature identification functionality including at least one of: N-gram identification functionality and co-occurrence identification functionality, and at least one of: skip-gram identification functionality, switch-gram identification functionality and previously used by user feature identification functionality.

There is yet further provided in accordance with another preferred embodiment of the present invention a computer-assisted language correction system including a grammatical error suspector evaluating at least most of the words in an language input on the basis of their fit within a context of the language input and a correction generator operative to provide a correction output based at least partially on an evaluation performed by the suspector.

Preferably, the computer-assisted language correction system also includes an alternatives generator, generating on the basis of the language input, a text-based representation providing multiple alternatives for at least one of the at least most words in the language input, and a selector for selecting among at least the multiple alternatives for each of the at least one of the at least most words in the language input, and the correction generator is operative to provide the correction output based on selections made by the selector.

In accordance with a preferred embodiment of the present invention the computer-assisted language correction system also includes a suspect word output indicator indicating an extent to which at least some of the at least most of the words in the language input is suspect as containing grammatical error.

Preferably, the correction generator includes an automatic corrected language generator operative to provide a corrected text output based at least partially on an evaluation performed by the suspector, without requiring user intervention.

There is also provided in accordance with still another preferred embodiment of the present invention a computer-assisted language correction system including a grammatical error suspector evaluating words in an language input, an alternatives generator, generating multiple alternatives for at least some of the words in the language input evaluated as suspect words by the suspector, at least one of the multiple alternatives for a word in the language input being consistent with a contextual

feature of the word in the language input, a selector for selecting among at least the multiple alternatives and a correction generator operative to provide a correction output based at least partially on a selection made by the selector.

There is further provided in accordance with yet another preferred embodiment of the present invention a computer-assisted language correction system including a grammatical error suspector evaluating words in an language input and identifying suspect words, an alternatives generator, generating multiple alternatives for the suspect words, a selector, grading each the suspect word as well as ones of the multiple alternatives therefor generated by the alternatives generator according to multiple selection criteria, and applying a bias in favor of the suspect word vis-à-vis ones of the multiple alternatives therefor generated by the alternatives generator, and a correction generator operative to provide a correction output based at least partially on a selection made by the selector.

Preferably, the correction generator includes a corrected language input generator operative to provide a corrected language output based on selections made by the selector without requiring user intervention.

There is even further provided in accordance with still another preferred embodiment of the present invention a computer-assisted language correction system including context based scoring of various alternative corrections, based at least partially on contextual feature-sequence (CFS) frequencies of occurrences in an internet corpus.

Preferably, the computer-assisted language correction system also includes at least one of spelling correction functionality, misused word correction functionality, grammar correction functionality and vocabulary enhancement functionality, cooperating with the context based scoring.

In accordance with a preferred embodiment of the present invention the context based scoring is also based at least partially on normalized CFS frequencies of occurrences in an internet corpus. Additionally or alternatively, the context based scoring is also based at least partially on a CFS importance score. Additionally, the CFS importance score is a function of at least one of the following: operation of a part-of-

speech tagging and sentence parsing functionality; a CFS length; a frequency of occurrence of each of the words in the CFS and a CFS type.

There is still further provided in accordance with yet another preferred embodiment of the present invention a computer-assisted language correction system including vocabulary enhancement functionality including vocabulary-challenged words identification functionality, alternative vocabulary enhancements generation functionality and context based scoring functionality, based at least partially on contextual feature-sequence (CFS) frequencies of occurrences in an internet corpus, the alternative vocabulary enhancements generation functionality including thesaurus pre-processing functionality operative to generate candidates for vocabulary enhancement.

There is also provided in accordance with yet another preferred embodiment of the present invention a computer-assisted language correction system including an alternatives generator, generating on the basis of an input sentence a text-based representation providing multiple alternatives for each of a plurality of words in the sentence, a selector for selecting among at least the multiple alternatives for each of the plurality of words in the sentence, a confidence level assigner operative to assign a confidence level to the selected alternative from said multiple alternatives and a correction generator operative to provide a correction output based on selections made by the selector and at least partially on the confidence level.

Preferably, the multiple alternatives are evaluated based on contextual feature sequences (CFSs) and the confidence level is based on at least one of the following parameters: number, type and scoring of selected CFSs, a measure of statistical significance of frequency of occurrence of the multiple alternatives, in the context of the CFSs, degree of consensus on the selection of one of the multiple alternatives, based on preference metrics of each of the CFSs and word similarity scores of the multiple alternatives, a non-contextual similarity score of the one of the multiple alternatives being above a first predetermined minimum threshold and an extent of contextual data available, as indicated by the number of the CFSs having CFS scores above a second predetermined minimum threshold and having preference scores over a third predetermined threshold.

There is also provided in accordance with yet another preferred embodiment of the present invention a computer-assisted language correction system including a punctuation error suspector evaluating at least some of the words and punctuation in a language input on the basis of their fit within a context of the language input based on frequency of occurrence of feature-grams of the language input in an internet corpus and a correction generator operative to provide a correction output based at least partially on an evaluation performed by the suspector.

Preferably, the correction generator includes at least one of missing punctuation correction functionality, superfluous punctuation correction functionality and punctuation replacement correction functionality.

There is further provided in accordance with still another preferred embodiment of the present invention a computer-assisted language correction system including a grammatical element error suspector evaluating at least some of the words in a language input on the basis of their fit within a context of the language input based on frequency of occurrence of feature-grams of the language input in an internet corpus and a correction generator operative to provide a correction output based at least partially on an evaluation performed by the suspector.

Preferably, the correction generator includes at least one of missing grammatical element correction functionality, superfluous grammatical element correction functionality and grammatical element replacement correction functionality. Additionally or alternatively, the grammatical element is one of an article, a preposition and a conjunction.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be understood and appreciated more fully from the following detailed description, taken in conjunction with the drawings in which:

Fig. 1 is a simplified block diagram illustration of a system and functionality for computer-assisted language correction constructed and operative in accordance with a preferred embodiment of the present invention;

Fig. 2 is a simplified flow chart illustrating spelling correction functionality, preferably employed in the system and functionality of Fig. 1;

Fig. 3 is a simplified flow chart illustrating misused word and grammar correction functionality, preferably employed in the system and functionality of Fig. 1;

Fig. 4 is a simplified flow chart illustrating vocabulary enhancement functionality, preferably employed in the system and functionality of Fig. 1;

Fig. 5 is a simplified block diagram illustrating contextual-feature-sequence (CFS) functionality, preferably employed in the system and functionality of Fig. 1;

Fig. 6A is a simplified flow chart illustrating spelling correction functionality forming part of the functionality of Fig. 2 in accordance with a preferred embodiment of the present invention;

Fig. 6B is a simplified flow chart illustrating misused word and grammar correction functionality forming part of the functionality of Fig. 3 in accordance with a preferred embodiment of the present invention;

Fig. 6C is a simplified flow chart illustrating vocabulary enhancement functionality forming part of the functionality of Fig. 4 in accordance with a preferred embodiment of the present invention;

Fig. 7A is a simplified flow chart illustrating functionality for generating alternative corrections which is useful in the functionalities of Figs. 2 and 3;

Fig. 7B is a simplified flow chart illustrating functionality for generating alternative enhancements which is useful in the functionality of Fig. 4;

Fig. 8 is a simplified flow chart illustrating functionality for non-contextual word similarity-based scoring and contextual scoring, preferably using an internet corpus, of various alternative corrections useful in the spelling correction functionality of Fig. 2;

Fig. 9 is a simplified flow chart illustrating functionality for non-contextual word similarity-based scoring and contextual scoring, preferably using an internet corpus, of various alternative corrections useful in the misused word and grammar correction functionalities of Figs. 3, 10 and 11 and in the vocabulary enhancement functionality of Fig. 4;

Fig. 10 is a simplified flowchart illustrating the operation of missing article, preposition and punctuation correction functionality; and

Fig. 11 is a simplified flowchart illustrating the operation of superfluous article, preposition and punctuation correction functionality.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Reference is now made to Fig. 1, which is a simplified block diagram illustration of a system and functionality for computer-assisted language correction constructed and operative in accordance with a preferred embodiment of the present invention. As seen in Fig. 1, text for correction is supplied to a language correction module 100 from one or more sources, including, without limitation, word processor functionality 102, machine translation functionality 104, speech-to-text conversion functionality 106, optical character recognition functionality 108 and any other text source 110, such as instant messaging or the internet.

Language correction module 100 preferably includes spelling correction functionality 112, misused word and grammar correction functionality 114 and vocabulary enhancement functionality 116.

It is a particular feature of the present invention that spelling correction functionality 112, misused word and grammar correction functionality 114 and vocabulary enhancement functionality 116 each interact with contextual-feature-sequence (CFS) functionality 118, which utilizes an internet corpus 120.

A contextual-feature-sequence or CFS is defined for the purposes of the present description as including, N-grams, skip-grams, switch-grams, co-occurrences, "previously used by user features" and combinations thereof, which are in turn defined hereinbelow with reference to Fig. 5. It is noted that for simplicity and clarity of description, most of the examples which follow employ n-grams only. It is understood that the invention is not so limited.

The use of an internet corpus is important in that it provides significant statistical data for an extremely large number of contextual-feature-sequences, resulting in highly robust language correction functionality. In practice, combinations of over two words have very poor statistics in conventional non-internet corpuses but have acceptable or good statistics in internet corpuses.

An internet corpus is a large representative sample of natural language text which is collected from the world wide web, usually by crawling on the internet and

collecting text from website pages. Preferably, dynamic text, such as chat transcripts, texts from web forums and texts from blogs, is also collected. The collected text is used for accumulating statistics on natural language text. The size of an internet corpus can be, for example, one trillion (1,000,000,000,000) words or several trillion words, as opposed to more typical corpus sizes of up to 2 billion words. A small sample of the web, such as the web corpus, includes 10 billion words, which is significantly less than one percent of the web texts indexed by search engines, such as GOOGLE®. The present invention can work with a sample of the web, such as the web corpus, but preferably it utilizes a significantly larger sample of the web for the task of text correction.

An internet corpus is preferably employed in one of the following two ways:

One or more internet search engines is employed using a CFS as a search query. The number of results for each such query provides the frequency of occurrence of that CFS.

A local index is built up over time by crawling and indexing the internet. The number of occurrences of each CFS provides the CFS frequency. The local index, as well as the search queries, may be based on selectable parts of the internet and may be identified with those selected parts. Similarly, parts of the internet may be excluded or appropriately weighted in order to correct anomalies between internet usage and general language usage. In such a way, websites that are reliable in terms of language usage, such as news and government websites, may be given greater weight than other websites, such as chat or user forums.

Preferably, input text is initially supplied to spelling correction functionality 112 and thereafter to misused word and grammar correction functionality 114. The input text may be any suitable text and in the context of word processing is preferably a part of a document, such as a sentence. Vocabulary enhancement functionality 116 preferably is operated at the option of a user on text that has already been supplied to spelling correction functionality 112 and to misused word and grammar correction functionality 114.

Preferably, the language correction module 100 provides an output which includes corrected text accompanied by one or more suggested alternatives for each corrected word or group of words.

Reference is now made to Fig. 2, which is a simplified flow chart illustrating spelling correction functionality, preferably employed in the system and functionality of Fig. 1. As seen in Fig. 2, the spelling correction functionality preferably comprises the following steps:

identifying spelling errors in an input text, preferably using a conventional dictionary enriched with proper names and words commonly used on the internet;

grouping spelling errors into clusters, which may include single or multiple words, consecutive or near consecutive, having spelling mistakes and selecting a cluster for correction. This selection attempts to find the cluster which contains the largest amount of correct contextual data. Preferably, the cluster that has the longest sequence or sequences of correctly spelled words in its vicinity is selected. The foregoing steps are described hereinbelow in greater detail with reference to Fig. 6A.

generating one or preferably more alternative corrections for each cluster, preferably based on an algorithm described hereinbelow with reference to Fig. 7A;

at least partially non-contextual word similarity-based scoring and contextual scoring, preferably using an internet corpus, of the various alternative corrections, preferably based on a spelling correction alternatives scoring algorithm, described hereinbelow with reference to Fig. 8;

for each cluster, selection of a single spelling correction and presentation of most preferred alternative spelling corrections based on the aforesaid scoring; and

providing a corrected text output incorporating the single spelling correction for each misspelled cluster, which replaces a misspelled cluster.

The operation of the functionality of Fig. 2 may be better understood from a consideration of the following example:

The following input text is received:

Physical ecudation can assits in strenghing muscles. Some students should eksersiv daily to inprove their strenth and helth becals thay ea so fate.

The following words are identified as spelling errors:

ecudation, assits; strenghing; eksersiv; inprove; strenth; helth; becals; thay, ea.

It is noted that "fate" is not identified as a spelling error inasmuch as it appears in a dictionary.

The following clusters are selected, as seen in Table 1:

TABLE 1

CLUSTER #	CLUSTER
1	eksersiv
2	inprove their strenth
3	ecudation
4	assits in strenghing
5	helth becals thay ea

Regarding cluster 2, it is noted that "their" is correctly spelled, but nevertheless included in a cluster since it is surrounded by misspelled words.

Cluster 1, "eksersiv" is selected for correction inasmuch as it has the longest sequence or sequences of correctly spelled words in its vicinity.

The following alternative corrections are generated for the misspelled word "eksersiv":

excessive, expressive, obsessive, assertive, exercise, extensive, exclusive, exertion, excised, exorcism.

Each alternative correction is given a non-contextual word similarity score, based on the similarity of sound and character string to the misspelled word, for example, as seen in Table 2:

TABLE 2

ALTERNATIVE	NON CONTEXTUAL WORD SIMILARITY SCORE
excessive	0.90
expressive	0.83
exercise	0.80
exorcism	0.56

The non-contextual score may be derived in various ways. One example is by using the Levelnshtein Distance algorithm which is available on http://en.wikipedia.org/wiki/Levenshtein_distance. This algorithm can be implied on word strings, word phonetic representation, or a combination of both.

Each alternative is also given a contextual score, as seen in Table 3, based on its fit in the context of the input sentence. In this example, the context that is used is "Some students should <eksersiv> daily"

TABLE 3

ALTERNATIVE	CONTEXTUAL SCORE FOR "Some students should [alternative] daily"	NON CONTEXTUAL WORD SIMILARITY SCORE
Excessive	0.00	0.90
Expressive	0.00	0.83
Exercise	1.00	0.80
Exorcism	0.00	0.56
Excised	0.16	0.62

The contextual score is preferably derived as described hereinbelow with reference to Fig. 8 and is based on contextual feature sequence (CFS) frequencies in an internet corpus.

The word "exercise" is selected as the best alternative based on a combination of the contextual score and non-contextual word similarity score, as described hereinbelow with reference to Fig. 8.

All of the clusters are corrected in a similar manner. The spelling-corrected input text, following spelling correction in accordance with a preferred embodiment of the present invention is:

> Physical education can assist in strengthening muscles. Some students should exercise daily to improve their strength and health because they are so fate.

It is noted that there remains a misused word in the spelling-corrected input text. The word "fate" requires correction by the misused word and grammar correction algorithm described hereinbelow with reference to Fig. 3.

Reference is now made to Fig. 3, which is a simplified flow chart illustrating misused word and grammar correction functionality, preferably employed in the system and functionality of Fig. 1. The misused word and grammar correction functionality provides correction of words which are correctly spelled but misused in the context of the input text and correction of grammar mistakes, including use of a grammatically incorrect word in place of grammatically correct word, the use of a superfluous word and missing words and punctuation.

As seen in Fig. 3, the misused word and grammar correction functionality preferably comprises the following steps:

identifying suspected misused words and words having grammar mistakes in a spelling-corrected input text output from the spelling correction functionality of Fig. 2, preferably by evaluating the fit of at least most of the words within the context of the input sentence;

grouping suspected misused words and words having grammar mistakes into clusters, which are preferably non-overlapping; and

selecting a cluster for correction. The identifying, grouping and selecting steps are preferably based on an algorithm described hereinbelow with reference to Fig. 6B.

generating one or preferably more alternative corrections for each cluster, preferably based on an alternative correction generation algorithm described hereinbelow with reference to Fig. 7A;

generating one or preferably more alternative corrections for each cluster, based on a missing article, preposition and punctuation correction algorithm described hereinbelow with reference to Fig. 10;

generating one or preferably more alternative corrections for each cluster, based on a superfluous article, preposition and punctuation correction algorithm described hereinbelow with reference to Fig. 11;

at least partially context-based and word similarity-based scoring of the various alternative corrections, preferably based on a misused word and grammar correction alternatives scoring algorithm, described hereinbelow with reference to Fig. 9;

for each cluster, selection of a single misused word and grammar correction and presentation of most preferred alternative misused word and grammar corrections based on the aforesaid scoring as also described hereinbelow with reference to Fig. 9; and

providing a spelling, misused word and grammar-corrected text output incorporating the single misused word and grammar correction for each cluster, which replaces an incorrect cluster.

Preferably, the scoring includes applying a bias in favor of the suspect word vis-à-vis ones of the multiple alternatives therefor, the bias being a function of an input uncertainty metric indicating uncertainty of a person providing the input.

The operation of the functionality of Fig. 3 may be better understood from a consideration of the following example:

The following input text is received:

I have money book

The following words are identified as suspected misused words:

money, book

The following cluster is generated:

money book

The following are examples of alternative corrections which are generated for the cluster (partial list):

money books; money back; money box; money bulk; money Buick; money ebook; money bank; mini book; mummy book; Monet book; honey book; mannerly book; mono book; Monday book; many books; mini bike; mummy back; monkey bunk; Monday booked; Monarchy back; Mourned brook

The results of at least partially contextual scoring using an internet corpus context-based and non-contextual word similarity-based scoring are presented in Table 4:

TABLE 4

CLUSTER	NON CONTEXTUAL SIMILARITY SCORE	CONTEXTUAL SCORE	GLOBAL SCORE
money back	0.72	0.30	0.216
many books	0.84	1.00	0.840
mini bike	0.47	0.75	0.352
money box	0.79	0.40	0.316
money bank	0.65	0.50	0.325
Monday booked	0.70	0.50	0.350
monkey bunk	0.54	0.00	0.000

It is appreciated that there exist various ways of arriving at a global score. The preferred global score is based on the algorithm described hereinbelow with reference to Fig. 9.

Based on the above scoring the alternative "many books" is selected. Thus, the corrected text is:

I have many books.

Reference is now made to Fig. 4, which is a simplified flow chart illustrating vocabulary enhancement functionality, employed in the system and functionality of Fig. 1. As seen in Fig. 4, the vocabulary enhancement functionality preferably comprises the following steps:

identifying vocabulary-challenged words having suspected suboptimal vocabulary usage in a spelling, misused word and grammar-corrected input text output from the misused word and grammar correction functionality of Fig. 3;

grouping vocabulary-challenged words into clusters, which are preferably non-overlapping;

selecting a cluster for correction. The identifying, grouping and selecting steps are preferably based on an algorithm described hereinbelow with reference to Fig. 6C.

generating one or preferably more alternative vocabulary enhancements for each cluster, preferably based on a vocabulary enhancement generation algorithm, described hereinbelow with reference to Fig. 7B;

non-contextual word similarity-based scoring and contextual scoring, preferably using an internet corpus, of the various alternative vocabulary enhancements, preferably based on a vocabulary enhancement alternatives scoring algorithm, described hereinbelow with reference to Fig. 9;

for each cluster, selection of a single vocabulary enhancement and presentation of most preferred alternative vocabulary enhancements based on the aforesaid scoring, also described hereinbelow with reference to Fig. 9; and

providing vocabulary enhancement proposals for each sub-optimal vocabulary cluster.

The operation of the functionality of Fig. 4 may be better understood from a consideration of the following example:

The following spelling, misused word and grammar-connected input text is provided:

Wearing colorful clothes will separate us from the rest of the children in the school.

Using the functionality described hereinbelow with reference to Fig. 6C, the following cluster is selected for vocabulary enhancement:

separate

Using the functionality described hereinbelow with reference to Fig. 7B, the following alternative cluster corrections, seen in Table 5, are generated based on the preprocessed lexical database that is described in Fig. 7B (partial list):

TABLE 5

WORD	FREQUENCY OF OCCURRENCE IN INTERNET CORPUS	PREPROCESSED RICHNESS SCORE
Furcated	1500	1.00
Disjoin	7800	0.97
Classify	1530500	0.93
Discriminate	1600400	0.91
Differentiate	1805600	0.90
Divide	4605800	0.88
Distinguish	4640900	0.88
Split	14003700	0.75
Branch	17106200	0.75
Separate	32765953	0.73
Break	39643000	0.72
Part	211805500	0.66

The various alternative vocabulary enhancements are scored using an internet corpus based at least partially on their fit within the context of the input text and also on their word meaning similarity to the vocabulary-challenged word "separate".

Using the functionality described hereinbelow with reference to Fig. 5, the following CFSs are generated (partial list):

'will separate', 'separate us', 'clothes will separate', 'will separate us', 'separate us from'

Using the functionality described hereinbelow with reference to Stage IIA of Fig. 9, the matrix of frequencies of occurrence in an internet corpus seen in Table

6 is generated for the partial list of alternative cluster corrections in the above list of CFSs:

TABLE 6

CFS/ ALTERNATIVE CLUSTER CORRECTIONS	will separate	separate us	clothes will separate	will separate us	separate us from
Classify	18180	1270	0	0	0
Discriminate	6440	215	0	0	0
Differentiate	24420	8365	0	740	5555

All CFSs for which all alternative corrections have a zero frequency of occurrence are eliminated. In this example the following feature-gram is eliminated:

'clothes will separate'

Thereafter, all CFSs which are entirely included in other CFSs having at least a minimum threshold frequency of occurrence are eliminated. For example, the following feature-grams are eliminated:

'will separate', 'separate us'

In this example, the remaining CFSs are the feature-grams:

'will separate us', 'separate us from'

Using the final preference score described hereinbelow in stages IID and IIE with reference to Fig. 9, the alternative "differentiate" is selected and the enhanced sentence is:

Wearing colorful clothes will differentiate us from the rest of the children in the school.

Reference is now made to Fig. 5, which is a simplified block diagram illustrating contextual-feature-sequence (CFS) functionality 118 (Fig. 1) useful in the system and functionality for computer-assisted language correction of a preferred embodiment of the present invention.

The CFS functionality 118 preferably includes feature extraction functionality including N-gram extraction functionality and optionally at least one of

skip-gram extraction functionality; switch-gram extraction functionality; co-occurrence extraction functionality; and previously used by user feature extraction functionality.

The term N-gram, which is a known term of the art, refers to a sequence of N consecutive words in an input text. The N-gram extraction functionality may employ conventional part-of-speech tagging and sentence parsing functionality in order to avoid generating certain N-grams which, based on grammatical considerations, are not expected to appear with high frequency in a corpus, preferably an internet corpus.

For the purposes of the present description, the term "skip-gram extraction functionality" means functionality operative to extract "skip-grams" which are modified n-grams which leave out certain non-essential words or phrases, such as adjectives, adverbs, adjectival phrases and adverbial phrases, or which contain only words having predetermined grammatical relationships, such as subject-verb, verb-object, adverb-verb or verb-time phrase. The skip-gram extraction functionality may employ conventional part-of-speech tagging and sentence parsing functionality to assist in deciding which words may be skipped in a given context.

For the purposes of the present description, the term "switch-gram extraction functionality" means functionality which identifies "switch grams", which are modified n-grams in which the order of appearance of certain words is switched. The switch-gram extraction functionality may employ conventional part-of-speech tagging and sentence parsing functionality to assist in deciding which words may have their order of appearance switched in a given context.

For the purposes of the present description, the term "co-occurrence extraction functionality" means functionality which identifies word combinations in an input sentence or an input document containing many input sentences, having input text word co-occurrence for all words in the input text other than those included in the N-grams, switch-grams or skip-grams, together with indications of distance from an input word and direction, following filtering out of commonly occurring words, such as prepositions, articles, conjunctions and other words whose function is primarily grammatical.

For the purposes of the present description, the term "previously used by user feature extraction functionality" means functionality which identifies words used

by a user in other documents, following filtering out of commonly occurring words, such as prepositions, articles, conjunctions and other words whose function is primarily grammatical.

For the purposes of the present description, N-grams, skip-grams, switch-grams and combinations thereof are termed feature-grams.

For the purposes of the present description, N-grams, skip-grams, switch-grams, co-occurrences, "previously used by user features" and combinations thereof are termed contextual-feature-sequences or CFSs.

The functionality of Fig. 5 preferably operates on individual words or clusters of words in an input text.

The operation of the functionality of Fig. 5 may be better understood from a consideration of the following example:

The following input text is provided:

Cherlock Homes the lead character and chief inspecter has been cold in by the family doctor Dr Mortimer , to invesigate the death of sir Charles"

For the cluster "Cherlock Homes" in the input text, the following CFSs are generated:

N-grams:

2-grams: Cherlock Homes; Homes the

3-grams: Cherlock Homes the; Homes the lead

4-grams: Cherlock Homes the lead; Homes the lead character

5-grams: Cherlock Homes the lead character

Skip-grams:

Cherlock Homes the character; Cherlock Homes the chief inspecter; Cherlock Homes the inspecter; Cherlock Homes has been cold

Switch gram:

The lead character Cherlock Homes

Co-occurrences in input text:

Character; inspector; investigate; death

Co-occurrences in document containing the input text:

Arthur Conan Doyle; story

Co-occurrence in other documents of user:

mystery

For the cluster "cold" in the input text, the following CFSs are generated:

N-grams:

2-grams: been cold; cold in

3-grams: has been cold; been cold in; cold in by

4-grams: inspector has been cold; has been cold in; been cold in by; cold in by the

5-grams: chief inspector has been cold; inspector has been cold in; has been cold in by; been cold in by the; cold in by the family

Skip-grams:

cold in to investigate; Cherlock has been cold; cold by the doctor; cold by Dr Mortimer; character has been cold

The CFSs are each given an "importance score" based on at least one of, preferably more than one of and most preferably all of the following:

a. operation of conventional part-of-speech tagging and sentence parsing functionality. A CFS which includes parts of multiple parsing tree nodes is given a relatively low score. The larger the number of parsing tree nodes included in a CFS, the lower is the score of that CFS.

b. length of the CFS. The longer the CFS, the higher the score.

c. frequency of occurrence of each of the words in the CFS other than the input word. The higher the frequency of occurrence of such words, the lower the score.

d. type of CFS. For example, an N-gram is preferred over a co-occurrence. A co-occurrence in an input sentence is preferred over a co-occurrence in an input document and a co-occurrence in an input document is preferred over "previously used by user features".

Referring to the above example, typical scores are as seen in Table 7:

TABLE 7

CFS TYPE	CFS	SCORE
N-gram: 2-gram	Cherlock Homes	0.50
N-gram: 2-gram	Homes the	0.30
N-gram: 3-gram	Cherlock Homes the	0.70
N-gram: 3-gram	Homes the lead	0.70
N-gram: 4-gram	Cherlock Homes the lead	0.90
N-gram: 4-gram	Homes the lead character	0.90
N-gram: 5-gram	Cherlock Homes the lead character	1.00
Skip-gram	Cherlock Homes the character	0.80
Skip-gram	Cherlock Homes the chief inspecter	0.95
Skip-gram	Cherlock Homes the inspecter	0.93
Skip-gram	Cherlock Homes has been cold	0.93
Switch gram	The lead character Cherlock Homes	0.95
Co-occurrence in input text	Character	0.40
Co-occurrence in input text	Inspector	0.40
Co-occurrence in input text	Investigate	0.40
Co-occurrence in input text	Death	0.40
Co-occurrence in document containing the input text:	Arthur Conan Doyle	0.50
Co-occurrence in document containing the input text:	Story	0.30
Co-occurrence in other documents of user	Mystery	0.20

These CFSs and their importance scores are used in the functionality described hereinbelow with reference to Figs. 8 & 9 for context based scoring of various

alternative cluster corrections, based on the CFS frequencies of occurrences in an internet corpus.

Reference is now made to Fig. 6A, which is a simplified flow chart illustrating functionality for identifying misspelled words in the input text; grouping misspelled words into clusters, which are preferably non-overlapping; and selecting a cluster for correction.

As seen in Fig. 6A, identifying misspelled words is preferably carried out by using a conventional dictionary enriched with proper names and words commonly used on the internet.

Grouping misspelled words into clusters is preferably carried out by grouping consecutive or nearly consecutive misspelled words into a single cluster along with misspelled words which have a grammatical relationship.

Selecting a cluster for correction is preferably carried out by attempting to find the cluster which contains the largest amount of non-suspected contextual data. Preferably, the cluster that has the longest sequence or sequences of correctly spelled words in its vicinity is selected.

Reference is now made to Fig. 6B, which is a simplified flow chart illustrating functionality for identifying suspected misused words and words having grammar mistakes in a spelling-corrected input text; grouping suspected misused words and words having grammar mistakes into clusters, which are preferably non-overlapping; and selecting a cluster for correction.

Identifying suspected misused words is preferably carried out as follows:

feature-grams are generated for each word in the spelling-corrected input text;

the frequency of occurrence of each of the feature-grams in a corpus, preferably an internet corpus, is noted;

the number of suspected feature-grams for each word is noted. Suspected feature-grams have a frequency which is significantly lower than their expected frequency or which lies below a minimum frequency threshold. The expected frequency of a feature-gram is estimated on the basis of the frequencies of its constituent elements and combinations thereof.

a word is suspected if the number of suspected feature-grams containing the word exceeds a predetermined threshold.

In accordance with a preferred embodiment of the invention, the frequency of occurrence of each feature-gram in the spelling-corrected input text in a corpus (FREQ F-G), preferably an internet corpus, is ascertained. The frequency of occurrence of each word in the spelling-corrected input text in that corpus (FREQ W) is also ascertained and the frequency of occurrence of each feature-gram without that word (FREQ FG-W) is additionally ascertained.

An expected frequency of occurrence of each feature-gram (EFREQ F-G) is calculated as follows:

EFREQ F-G = FREQ F-G-W * FREQ W/(TOTAL OF FREQUENCIES OF ALL WORDS IN THE CORPUS)

If the ratio of the frequency of occurrence of each feature-gram in the spelling-corrected input text in a corpus, preferably an internet corpus, to the expected frequency of occurrence of each feature-gram, FREQ F-G/EFREQ F-G, is less than a predetermined threshold, or if FREQ F-G is less than another predetermined threshold, the feature-gram is considered to be a suspected feature-gram. Every word that is included in a suspected feature-gram is considered to be a suspected misused word or a word having a suspected grammar mistake.

The operation of the functionality of Fig. 6B for identifying suspected misused words and words having grammar mistakes in a spelling-corrected input text may be better understood from a consideration of the following example:

The following spelling-corrected input text is provided:

I have money book

The feature-grams include the following:

I; I have; I have money; I have money book

Table 8 indicates the frequencies of occurrence in an internet corpus of the above feature-grams:

TABLE 8

WORD / FREQUENCY	1-GRAM	2-GRAMS	3-GRAMS	4-GRAMS
I	I 2744650600	I have 154980000	I have money 24600	I have money book 0
Have	have 1500601100	have money 390300	have money book 0	
Money	money 124877160	money book 3100		
Book	book 164065000			

The expected frequencies of occurrence are calculated for each of the 2-grams as follows:

EFREQ F-G = (FREQ F-G-W * FREQ W)/(TOTAL OF FREQUENCIES OF ALL WORDS IN THE CORPUS)

For example, for a 2-gram,

the expected 2-gram frequency for a 2-gram (x,y) = (1-gram frequency of x * 1-gram frequency of y)/Number of words in the internet corpus. e.g., Trillion (1,000,000,000,000) words.

The ratio of the frequency of occurrence of each feature-gram in the spelling-corrected input text in a corpus, preferably an internet corpus, to the expected frequency of occurrence of each feature-gram is calculated as follows:

FREQ F-G/EFREQ F-G

The ratio of the frequency of occurrence of each of the above 2-grams in the spelling-corrected input text in a corpus, preferably an internet corpus, to the expected frequency of occurrence of each of the above 2-grams are seen in Table 9:

TABLE 9

2-GRAMS	FREQ F-G	EFREQ F-G	FREQ F-G/EFREQ F-G
I have	154980000	4118625.7	37.60
have money	390300	187390.8	2.08
money book	3100	20487.9	0.15

It is seen that FREQ F-G of "money book" is substantially lower than its expected frequency and thus FREQ F-G/EFREQ F-G may be considered to be lower than a predetermined threshold, such as 1, and therefore the cluster "money book" is suspected.

It is seen that the 3-gram and the 4-gram including the words "money book" both have a zero frequency in the internet corpus. This can also be a basis for considering "money book" to be suspect.

Grouping suspected misused words and words having grammar mistakes into clusters is preferably carried out as follows: consecutive or nearly consecutive suspected misused words are grouped into a single cluster; and suspected misused words which have a grammatical relationship between themselves are grouped into the same cluster.

Selecting a cluster for correction is preferably carried out by attempting to find the cluster which contains the largest amount of non-suspected contextual data. Preferably, the cluster that has the longest sequence or sequences of non-suspected words in its vicinity is selected.

Reference is now made to Fig. 6C, which is a simplified flow chart illustrating functionality for identifying vocabulary-challenged words having suspected suboptimal vocabulary usage in a spelling, misused word and grammar-corrected input text; grouping vocabulary-challenged words into clusters, which are preferably non-overlapping; and selecting a cluster for correction.

Identifying vocabulary-challenged words is preferably carried out as follows:

pre-processing a thesaurus in order to assign language richness scores to each word which indicate the level of the word in a hierarchy wherein written language is preferred over spoken language; and wherein among internet sources, articles and books are preferred over chat and forums, for example, and wherein less frequently used words are preferred over more frequently used words;

further pre-processing of the thesaurus to eliminate words which are not likely candidates for vocabulary enhancement based on the results of the preceding pre-processing step and on grammatical rules;

additional pre-processing to indicate for each remaining word, candidates for vocabulary enhancement which have a language richness score higher than that of the input word; and

checking whether each word in the spelling, misused word and grammar-corrected input text appears as a remaining word in the multiple pre-processed thesaurus and identifying each such word which appears as a remaining word as a candidate for vocabulary enhancement.

Grouping vocabulary-challenged words into clusters, which are preferably non-overlapping, is optional and is preferably carried out as follows:

consecutive vocabulary-challenged words are grouped into a single cluster; and

vocabulary-challenged words which have a grammatical relationship are grouped into the same cluster.

Selecting a cluster for correction is preferably carried out by attempting to find the cluster which contains the largest amount of non vocabulary-challenged words. Preferably, the cluster that has the longest sequence or sequences of non vocabulary-challenged words in its vicinity is selected.

Reference is now made to Fig. 7A, which is a simplified flow chart illustrating functionality for generating alternative corrections for a cluster, which is useful in the functionalities of Figs. 2 and 3.

If the original input word is correctly spelled, it is considered as an alternative.

As seen in Fig. 7A, for each word in the cluster, a plurality of alternative corrections is initially generated in the following manner:

A plurality of words, taken from a dictionary, similar to each word in the cluster, both on the basis of their written appearance, expressed in character string similarity, and on the basis of sound or phonetic similarity, is retrieved. This functionality is known and available on the internet as freeware, such as GNU Aspell and Google ® GSpell. The retrieved and prioritized words provide a first plurality of alternative corrections. E.g., given the input word feezix, the word "physics" will be retrieved from the dictionary, based on a similar sound, even though it has only one character, namely "i", in common. The word "felix" will be retrieved, based on its string character similarity, even though it doesn't have a similar sound.

Additional alternatives may be generated by employing rules based on known alternative usages as well as accumulated user inputs. E.g., u → you, r → are, Im → I am.

Further alternatives may be generated based on grammatical rules, preferably employing pre-defined lists. A few examples follow:

singular/plural rules: If the input sentence is "**leaf** fall off trees in the autumn" the plural alternative "leaves" is generated.

article rules: If the input text is "**a** old lady", the alternative articles "an" & "the" are generated.

preposition rules: If the input text is "I am interested **of** football", the alternative prepositions "in", "at", "to", "on", "through", … are generated.

verb inflection rules: If the input text is "He **leave** the room", the alternative verb inflections "left", "leaves", "had left", ... are generated.

merged words and split words rules: If the input text is "get **alot** fitter", the alternative "a lot" is generated.

If the input text is "we have to **wat ch** out", the alternative "watch" is generated.

If the input text is "do many **sittups**", the alternative "sit ups" is generated.

It is a particular feature of a preferred embodiment of the present invention that contextual information, such as CFSs and more particularly feature-grams, is employed to generate alternative corrections and not only for scoring such "contextually retrieved" alternative corrections. Frequently occurring word combinations, such as CFSs and more particularly feature-grams, may be retrieved from an existing corpus, such as an internet corpus.

The following example illustrates this aspect of the present invention:

If the input sentence is: "The cat has **kts**", the word "kts" may not be sufficiently similar in sound or writing to the word "kittens" such that absent this aspect of the invention, "kittens" might not be one of the alternatives.

In accordance with this aspect of the present invention, by looking in an internet corpus for words which commonly appear after the n-gram "cat has", i.e., all words found as * in the query "cat has *", the following alternatives are retrieved:

nine lives; left; fleas; dandruff; kittens; tapeworms; adopted; retractile claws; been; urinated; diarrhea; eaten; swallowed; hairballs; gone; always been

In accordance with a preferred embodiment of the present invention, the "contextually retrieved" alternatives are then filtered, such that only contextually retrieved alternatives having some phonetic or writing similarity to the original word, in the present example "kts", remain. In this example, the alternative having the highest phonetic and writing similarity, "kittens", is retrieved.

Where the input text is generated automatically by an external system, such as an optical character recognition, speech-to-text or machine translation system, additional alternatives may be received directly from such system. Such additional alternatives typically are generated in the course of operation of such system. For example, in a machine translation system, the alternative translations of a word in a foreign language may be supplied to the present system for use as alternatives.

Once all of the alternatives for each of the words in the cluster have been generated, cluster alternatives for the entire cluster are generated by ascertaining all possible combinations of the various alternatives and subsequent filtering of the

combinations based on the frequency of their occurrence in a corpus, preferably an internet corpus.

The following example is illustrative:

If the input cluster is "money book", and the alternatives for the word "money" are:

Monday; many; monkey

and the alternatives for the word "book" are:

books; box; back

The following cluster alternatives are generated:

money books; money box; money back; Monday books; Monday box; Monday back; many books; many box; many back; monkey books; monkey box; monkey back; many book; monkey book; Monday book

Reference is now made to Fig. 7B, which is a simplified flow chart illustrating functionality for generating alternative enhancements for a cluster, which is useful in the functionality of Fig. 4.

As seen in Fig. 7B, a plurality of alternative enhancements is initially generated in the following manner:

If the original input word is correctly spelled, it is considered as an alternative.

A plurality of words, taken from a thesaurus or other lexical database, such as Princeton WordNet, freely available on the internet, which are lexically related to each word in the cluster, as synonyms, superset or subset, is retrieved. The retrieved and prioritized words provide a first plurality of alternative enhancements.

Additional alternatives may be generated by employing rules based on known alternative usages as well as accumulated user inputs.

It is a particular feature of a preferred embodiment of the present invention that contextual information, such as CFSs and more particularly feature-grams is employed to generate alternative enhancements and not only for scoring such "contextually retrieved" alternative enhancements. Frequently occurring word combinations, such as CFSs and more particularly feature-grams, may be retrieved from an existing corpus, such as an internet corpus.

Once all of the alternatives for each of the words in the cluster have been generated, alternatives for the entire cluster are generated by ascertaining all possible combinations of the various word alternatives and filtering of the combinations based on frequency of the occurrence of the resulting multi-word combinations in an existing corpus, such as the internet.

The following example is illustrative of the functionality of Fig. 7B:

The following input text is provided:

it was nice to meet you

The following alternatives for the word "nice" are generated by a lexical database, such as Princeton WordNet (partial list):

pleasant, good, kind, polite, fine, decent, gracious, courteous, considerate, enjoyable, agreeable, satisfying, delightful, entertaining, amiable, friendly, elegant, precise, careful, meticulous.

The following alternative for the word "nice" is generated by the application of predetermined rules:

cool

The following contextually retrieved alternatives for the word "nice" are generated, such as in response to a query such as "it was * to meet":

great; a pleasure; wonderful; lovely; good; impossible; fun; awesome; refreshing; exciting; agreed; fantastic; decided ; inspiring

Reference is now made to Fig. 8, which is a simplified flow chart illustrating functionality for context-based and word similarity-based scoring of various alternative enhancements useful in the spelling correction functionality of Fig. 2.

As seen in Fig. 8, the context-based and word similarity-based scoring of various alternative corrections proceeds in the following general stages:

I. NON-CONTEXTUAL SCORING - Various cluster alternatives are scored on the basis of similarity to a cluster in the input text in terms of their written appearance and sound similarity. This scoring does not take into account any contextual similarity outside of the given cluster.

II. CONTEXTUAL SCORING USING INTERNET CORPUS - Each of the various cluster alternatives is also scored on the basis of extracted contextual-

feature-sequences (CFSs), which are provided as described hereinabove with reference to Fig. 5. This scoring includes the following sub-stages:

IIA. Frequency of occurrence analysis is carried out, preferably using an internet corpus, on the various alternative cluster corrections produced by the functionality of Fig. 7A, in the context of the CFSs extracted as described hereinabove with reference to Fig. 5.

IIB. CFS selection and weighting of the various CFSs is carried out based on, inter alia, the results of the frequency of occurrence analysis of sub-stage IIA. Weighting is also based on relative inherent importance of various CFSs. It is appreciated that some of the CFSs may be given a weighting of zero and are thus not selected. The selected CFSs preferably are given relative weightings.

IIC. A frequency of occurrence metric is assigned to each alternative correction for each of the selected CFSs in sub-stage IIB.

IID. A reduced set of alternative cluster corrections is generated, based, inter alia, on the results of the frequency of occurrence analysis of sub-stage IIA, the frequency of occurrence metric of sub-stage IIC and the CFS selection and weighting of sub-stage IIB.

IIE. The cluster having the highest non-contextual similarity score in stage I is selected from the reduced set in sub-stage IID for use as a reference cluster correction.

IIF. A frequency of occurrence metric is assigned to the reference cluster correction of sub-stage IIE for each of the selected CFSs in stage IIB.

IIG. A ratio metric is assigned to each of the selected CFSs in sub-stage IIB which represents the ratio of the frequency of occurrence metric for each alternative correction for that feature to the frequency of occurrence metric assigned to the reference cluster of sub-stage IIE.

III. A most preferred alternative cluster correction is selected based on the results of stage I and the results of stage II.

IV. A confidence level score is assigned to the most preferred alternative cluster correction.

A more detailed description of the functionality described hereinabove in stages II - IV is presented hereinbelow:

With reference to sub-stage IIA, all of the CFSs which include the cluster to be corrected are generated as described hereinabove in Fig. 5. CFSs containing suspected errors, other than errors in the input cluster, are eliminated.

A matrix is generated indicating the frequency of occurrence in a corpus, preferably an internet corpus, of each of the alternative corrections for the cluster in each of the CFSs. All CFSs for which all alternative corrections have a zero frequency of occurrence are eliminated. Thereafter, all CFSs which are entirely included in other CFSs having at least a minimum threshold frequency of occurrence are eliminated.

The following example illustrates generation of a frequency of occurrence matrix:

The following input text is provided:

I lik tw play outside a lot

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

lik tw

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

like to; like two; lick two; lack two; lack true; like true

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'lik tw'; 'I lik tw'; 'lik tw play'; 'I lik tw play'; 'lik tw play outside'; 'I lik tw play outside'; 'lik tw play outside a'

Using the functionality described hereinabove with reference to Stage IIA, the matrix of frequencies of occurrence in an internet corpus seen in Table 10 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 10

CFS/ ALTERNATIVE CLUSTER CORRECTION	lik tw	I lik tw	lik tw play	I lik tw play	lik tw play outside	I lik tw play outside	lik tw play outside a
like to	75094300	3097250	432200	80500	1000	330	0
like two	327500	2400	0	0	0	0	0
lick two	600	0	0	0	0	0	0
lack two	1800	100	0	0	0	0	0
lack true	3400	90	0	0	0	0	0
like true	27500	350	0	0	0	0	0

All CFSs for which all alternative corrections have a zero frequency of occurrence are eliminated. In this example the following feature-gram is eliminated:

'lik tw play outside a'

Thereafter, all CFSs which are entirely included in other CFSs having at least a minimum threshold frequency of occurrence are eliminated. In this example the following feature-grams are eliminated:

'lik tw'; 'I lik tw'; 'lik tw play'; 'I lik tw play'; 'lik tw play outside'

In this example the only remaining CFS is the feature-gram:

'I lik tw play outside'.

The resulting matrix appears as seen in Table 11:

TABLE 11

CFS/ALTERNATIVE CLUSTER CORRECTIONS	**'I lik tw play outside'**
like to	330
like two	0
lick two	0
lack two	0
lack true	0
like true	0

The foregoing example illustrates the generation of a matrix in accordance with a preferred embodiment of the present invention. In this example, it is clear that "like to" is the preferred alternative correction. It is to be appreciated that in reality, the choices are not usually so straightforward. Accordingly, in further examples presented below, functionality is provided for making much more difficult choices among alternative corrections.

Returning to a consideration of sub-stage IIB, optionally, each of the remaining CFSs is given a score as described hereinabove with reference to Fig. 5. Additionally, CFSs which contain words introduced in an earlier correction iteration of the multi-word input and have a confidence level below a predetermined confidence level threshold are negatively biased.

In the general case, similarly to that described hereinabove in sub-stage IIC, preferably, a normalized frequency matrix is generated indicating the normalized frequency of occurrence of each CFS in the internet corpus. The normalized frequency matrix is normally generated from the frequency matrix by dividing each CFS frequency by a function of the frequencies of occurrence of the relevant cluster alternatives.

The normalization is operative to neutralize the effect of substantial differences in overall popularity of various alternative corrections. A suitable

normalization factor is based on the overall frequencies of occurrence of various alternative corrections in a corpus as a whole, without regard to particular CFSs.

The following example illustrates the generation of a normalized frequency of occurrence matrix:

The following input text is provided:

footprints of a mysterious haund said to be six feet tall

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

haund

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

hound; hand; sound

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'a mysterious haund'; 'haund said'

Using the functionality described hereinabove with reference to Stage IIC herein, the matrix of frequencies of occurrence and normalized frequencies of occurrence in an internet corpus seen in Table 12 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 12

	FREQUENCY			NORMALIZED FREQUENCY	
CFS/ ALTERNATIVE CLUSTER CORRECTION	**ALTERNATIVE CLUSTER CORRECTION**	**'a mysterious haund'**	**'haund said'**	**'a mysterious haund'**	**'haund said'**
Hound	542300	50	180	0.0000921	0.00033
Hand	95073600	700	5900	0.0000073	0.00006
Sound	59026500	590	600	0.0000099	0.00001

It may be appreciated from the foregoing example that words having the highest frequencies of occurrence may not necessarily have the highest normalized frequencies of occurrence, due to substantial differences in overall popularity of various alternative corrections. In the foregoing example, "hound" has the highest normalized frequencies of occurrence and it is clear from the context of the input text that "hound" is the correct word, rather than "hand" which has higher frequencies of occurrence in the internet corpus.

It is a particular feature of the present invention that normalized frequencies of occurrence, which neutralize substantial differences in overall popularity of various alternative corrections, are preferably used in selecting among the alternative corrections. It is appreciated that other metrics of frequency of occurrence, other than normalized frequencies of occurrence, may alternatively or additionally be employed as metrics. Where the frequencies of occurrence are relatively low or particularly high, additional or alternative metrics are beneficial.

It will be appreciated from the discussion that follows that additional functionalities are often useful in selecting among various alternative corrections. These functionalities are described hereinbelow.

In sub-stage IID, each alternative cluster correction which is less preferred than another alternative cluster correction according to both of the following metrics is eliminated:

i. having a word similarity score lower than the other alternative cluster correction; and

ii. having lower frequencies of occurrences and preferably also lower normalized frequencies of occurrence for all of the CFSs than the other alternative cluster correction.

The following example illustrates the elimination of alternative corrections as described hereinabove:

The following input text is provided:

I leav un a big house

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

leav un

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

leave in; live in; love in

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'I leav un a'; 'leav un a big'

Using the functionality described hereinabove with reference to Stage IIC herein, the matrix of frequencies of occurrence and normalized frequencies of occurrence in an internet corpus seen in Table 13 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 13

	FREQUENCY			**NORMALIZED FREQUENCY**	
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**CLUSTER ALTERNTIVE CORRECTION**	**'I leav un a'**	**'leav un a big'**	**'I leav un a'**	**'leav un a big'**
leave in	442650	1700	100	0.0038	0.00022
live in	15277750	266950	17800	0.0174	0.00116
love in	1023100	1880	290	0.0018	0.00028

In this example, the non-contextual similarity scores of the alternative cluster corrections are as indicated in Table 14:

TABLE 14

ALTERNATIVE CLUSTER CORRECTION	SIMILARITY SCORE
leave in	0.9
live in	0.8
love in	0.7

The alternative cluster correction "love in" is eliminated as it has a lower similarity score as well as lower frequencies of occurrence and lower normalized frequencies of occurrence than "live in". The alternative cluster correction "leave in" is not eliminated at this stage since its similarity score is higher than that of "live in".

As can be appreciated from the foregoing, the result of operation of the functionality of stage IID is a reduced frequency matrix and preferably also a reduced normalized frequency matrix, indicating the frequency of occurrence and preferably also the normalized frequency of occurrence of each of a reduced plurality of alternative corrections, each of which has a similarity score, for each of a reduced plurality of CFSs. The reduced set of alternative cluster corrections is preferably employed for all further alternative cluster selection functionalities as is seen from the examples which follow.

For each alternative correction in the reduced frequency matrix and preferably also in the reduced normalized frequency matrix, a final preference metric is generated. One or more of the following alternative metrics may be employed to generate a final preference score for each alternative correction:

The term "frequency function" is used below to refer to the frequency, the normalized frequency or a function of both the frequency and the normalized frequency.

A. One possible preference metric is the highest occurrence frequency function for each alternative cluster correction in the reduced matrix or matrices for any of the CFSs in the reduced matrix or matrices. For example, the various alternative cluster corrections would be scored as follows:

The following input text is provided:

A big agle in the sky

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

agle

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

ogle; eagle; angel

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'big agle'; 'agle in the sky'

Using the functionality described hereinabove with reference to Stage IIC herein, the matrix of frequencies of occurrence and normalized frequencies of occurrence in an internet corpus seen in Table 15 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 15

	FREQUENCY			NORMALIZED FREQUENCY	
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**ALTERNATIVE CLUSTER CORRECTION**	**'big agle'**	**'agle in the sky'**	**'big agle'**	**'agle in the sky'**
Ogle	95250	0	0	0	0
Eagle	3058320	1550	1200	0.00050	0.00039
Angel	5961100	780	1060	0.00013	0.00017

In this example, the non-contextual similarity scores of the alternative cluster corrections are as indicated in Table 16:

TABLE 16

ALTERNATIVE CLUSTER CORRECTION	SIMILARITY SCORE
ogle	0.97
eagle	0.91
angel	0.83

The alternative 'eagle' is selected because it has a CFS with a maximum frequency of occurrence.

B. Another possible preference metric is the average occurrence frequency function of all CFSs for each alternative correction. For example, the various alternative corrections would be scored as follows:

The following input text is provided:

A while ago sthe lived 3 dwarfs

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

sthe

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

the; they; she; there

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'ago sthe lived'; 'sthe lived 3'

Using the functionality described hereinabove with reference to Stage IIC herein, the matrix of frequencies of occurrence, normalized frequencies of occurrence and average frequency of occurrence in an internet corpus seen in Tables 17 and 18 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 17

	FREQUENCY		
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**ALTERNATIVE CLUSTER CORRECTION**	**'ago sthe lived'**	**'sthe lived 3'**
The	19401194700	0	0
They	702221530	300	45
She	234969160	215	65
there	478280320	3200	40

TABLE 18

	NORMALIZED FREQUENCY		AVERAGE
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**'ago sthe lived'**	**'sthe lived 3'**	**AVERAGE FREQUENCY OF OCCRRENCE**
The	0	0	0
they	0.0000004	0.00000006	172
She	0.0000009	0.00000027	140
there	0.0000066	0.00000008	1620

5

It is noted that "there" is selected based on the average frequency of occurrence.

In this example, the non-contextual similarity scores of the alternative cluster corrections are as indicated in Table 19:

TABLE 19

ALTERNATIVE CLUSTER CORRECTION	SIMILARITY SCORE
the	0.97
they	0.86
she	0.88
there	0.67

It is noted that the alternative cluster correction having the highest similarity score is not selected.

C. A further possible preference metric is the weighted sum, over all CFSs for each alternative correction, of the occurrence frequency function for each CFS multiplied by the score of that CFS as computed by the functionality described hereinabove with reference to Fig. 5.

D. A Specific Alternative Correction/CFS preference metric is generated, as described hereinabove with reference to sub-stages IIE - IIG, by any one or more, and more preferably most and most preferably all of the following operations on the alternative corrections in the reduced matrix or matrices:

i. The alternative cluster correction having the highest non-contextual similarity score is selected to be the reference cluster.

ii. A modified matrix is produced wherein in each preference matrix, the occurrence frequency function of each alternative correction in each feature gram is replaced by the ratio of the occurrence frequency function of each alternative correction to the occurrence frequency function of the reference cluster.

iii. A modified matrix of the type described hereinabove in ii. is further modified to replace the ratio in each preference metric by a function of the ratio which function reduces the computational importance of very large differences in ratios.

A suitable such function is a logarithmic function. The purpose of this operation is to de-emphasize the importance of large differences in frequencies of occurrence in the final preference scoring of the most preferred alternative corrections, while maintaining the importance of large differences in frequencies of occurrence in the final preference scoring, and thus elimination, of the least preferred alternative corrections.

iv. A modified matrix of the type described hereinabove in ii or iii is additionally modified by multiplying the applicable ratio or function of ratio in each preference metric by the appropriate CFS score. This provides emphasis based on correct grammatical usage and other factors which are reflected in the CFS score.

v. A modified matrix of the type described hereinabove in ii, iii or iv is additionally modified by generating a function of the applicable ratio, function of ratio, frequency of occurrence and normalized frequency of occurrence. A preferred function is generated by multiplying the applicable ratio or function of ratio in each preference metric by the frequency of occurrence of that CFS.

E. A final preference metric is computed for each alternative correction based on the Specific Alternative Correction/CFS preference metric as described hereinabove in D by multiplying the similarity score of the alternative correction by the sum of the Specific Alternative Correction/CFS preference metrics for all CFS for that Alternative Correction.

An example illustrating the use of such a modified matrix is as follows:

The following input text is provided:

I will be able to tach base with you next week

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

tach

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

teach; touch

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'able to tach'; 'to tach base'

Using the functionality described hereinabove with reference to sub-stages IIA & IIC hereinabove, the matrix of frequencies of occurrence and normalized frequencies of occurrence in an internet corpus seen in Table 20 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 20

	FREQUENCY			NORMALIZED FREQUENCY	
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**ALTERNATIVE CLUSTER CORRECTIONS**	**'able to tach'**	**'to tach base'**	**'able to tach'**	**'to tach base'**
Teach	15124750	103600	40	0.0068	0.000002
touch	23506900	45050	27150	0.0019	0.001154

It is noted that for one feature, both the frequency of occurrence and the normalized frequency of occurrence of "teach" are greater than those of "touch", but for another feature, both the frequency of occurrence and the normalized frequency of occurrence of "touch" are greater than those of "teach". In order to make a correct choice of an alternative correction, ratio metrics, described hereinabove with reference to sub-stage IIG, are preferably employed as described hereinbelow.

In this example, the non-contextual similarity scores of the alternative cluster corrections are as indicated in Table 21:

TABLE 21

ALTERNATIVE CLUSTER CORRECTION	**SIMILARITY SCORE**
teach	0.94
touch	0.89

It is seen that the reference cluster is "teach", since it has the highest similarity score. Nevertheless "touch" is selected based on the final preference score described hereinabove. This is not intuitive, as may be appreciated from a consideration of the above matrices which indicate that "teach" has the highest frequency of occurrence and the highest normalized frequency of occurrence. In this example, the final preference score indicates a selection of "touch" over "teach" since the ratio of frequencies of occurrence for a feature in which "touch" is favored is much greater than the ratio of frequencies of occurrence for the other feature in which "teach" is favored.

F. Optionally, an alternative correction may be filtered out on the basis of a comparison of frequency function values and preference metrics for that alternative correction and for the reference cluster using one or more of the following decision rules:

1. filtering out an alternative correction having a similarity score below a predetermined threshold and having a CFS frequency function that is less than the CFS frequency function of the reference cluster for at least one feature which has a CFS score which is higher than a predetermined threshold.

2. filtering out alternative corrections having a similarity score below a predetermined threshold and having a preference metric which is less than a predetermined threshold for at least one feature which has a CFS score which is higher than another predetermined threshold.

3. a. ascertaining the CFS score of each CFS;

b. for each CFS, ascertaining the CFS frequency functions for the reference cluster and for an alternative correction, thereby to ascertain whether the reference cluster or the alternative correction has a higher frequency function for that CFS;

c. summing the CFS scores of CFSs for which the alternative correction has a higher frequency than the reference cluster;

d. summing the CFS scores of CFSs for which the reference cluster has a higher frequency than the alternative correction; and

e. if the sum in c. is less than the sum in d. filtering out that alternative correction.

The following example illustrates the filtering functionality described above.

The following input text is provided:

I am faelling en love

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

faelling en

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

falling on; falling in; feeling on; feeling in

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'am faelling en'; 'faelling en love'; 'am faelling en love'; 'I am faelling en'

Using the functionality described hereinabove with reference to sub-stage IIA herein, the matrix of frequencies of occurrence in an internet corpus seen in Table 22 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 22

CFS/ALTERNATIVE CLUSTER CORRECTIONS	**'am faelling en'**	**'faelling en love'**	**'am faelling en love'**	**'I am faelling en'**
falling on	200	40	0	185
falling in	4055	341800	3625	3345
feeling on	435	70	0	370
feeling in	1035	1055	0	895

All CFSs which are entirely included in other CFSs having at least a minimum threshold frequency of occurrence are eliminated. For example the following feature-grams are eliminated:

'am faelling en'; 'faelling en love'

In this example the remaining CFSs are the feature-grams:

'am faelling en love'; 'I am faelling en'

In this example, the non-contextual similarity scores of the alternative cluster corrections are as indicated in Table 23:

TABLE 23

ALTERNATIVE CLUSTER CORRECTION	SIMILARITY SCORE
falling on	0.89
falling in	0.89
feeling on	0.82
feeling in	0.82

The alternative corrections "falling on", "feeling on" and "feeling in" are filtered out because they have zero frequency of occurrence for one of the CFSs.

G. As discussed hereinabove with reference to Stage III, a ranking is established based on the final preference metric developed as described hereinabove at A - E on the alternative corrections which survive the filtering in F. The alternative correction having the highest final preference score is selected.

H. As discussed hereinabove with reference to Stage IV, a confidence level is assigned to the selected alternative correction. This confidence level is calculated based on one or more of the following parameters:

a. number, type and scoring of selected CFSs as provided in sub-stage IIB above;

b. statistical significance of frequency of occurrence of the various alternative cluster corrections, in the context of the CFSs;

c. degree of consensus on the selection of an alternative correction, based on preference metrics of each of the CFSs and the word similarity scores of the various alternative corrections;

d. non-contextual similarity score (stage I) of the selected alternative cluster correction being above a predetermined minimum threshold.

e. extent of contextual data available, as indicated by the number of CFSs in the reduced matrix having CFS scores above a predetermined minimum threshold and having preference scores over another predetermined threshold.

If the confidence level is above a predetermined threshold, the selected alternative correction is implemented without user interaction. If the confidence level is below the predetermined threshold but above a lower predetermined threshold, the selected alternative correction is implemented but user interaction is invited. If the confidence level is below the lower predetermined threshold, user selection based on a prioritized list of alternative corrections is invited.

The following examples are illustrative of the use of confidence level scoring:

The following input text is provided:

He was not feeling wehl when he returned

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

wehl

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

wale; well

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'was not feeling wehl'; 'not feeling wehl when'; 'feeling wehl when he'; 'wehl when he returned'

Using the functionality described hereinabove with reference to sub-stage IIA herein, the matrix of frequencies of occurrence in an internet corpus seen in Table 24 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 24

CFS/ ALTERNATIVE CLUSTER CORRECTIONS	'was not feeling wehl'	'not feeling wehl when'	'feeling wehl when he'	'wehl when he returned'
Wale	0	0	0	0
Well	31500	520	100	140

The foregoing example illustrates that, according to all the criteria set forth in H above, the selection of 'well' over 'wale' has a high confidence level.

In the following example, the confidence level is somewhat less, due to the fact that the alternative correction 'back' has a higher frequency of occurrence than 'beach' in the CFS 'bech in the summer' but 'beach' has a higher frequency of occurrence than 'back' in the CFSs 'on the beech in' and 'the bech in the'. The alternative correction 'beach' is selected with an intermediate confidence level based on criterion H(c).

The following input text is provided:

I like to work on the bech in the summer

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

bech

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

beach; beech; back

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'on the bech in'; 'the bech in the'; 'bech in the summer'

Using the functionality described hereinabove with reference to sub-stage IIA, the matrix of frequencies of occurrence in an internet corpus seen in Table 25 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 25

CFS/ ALTERNATIVE CLUSTER CORRECTIONS	'on the bech in'	'the bech in the'	'bech in the summer'
Beach	110560	42970	2670
Beech	50	55	0
Back	15300	10390	20090

The alternative correction 'beach' is selected with an intermediate confidence level based on criterion H(c).

In the following example, the confidence level is even less, based on criterion H(a):

The following input text is received:

Exarts are restoring the British Museum's round reading room

Using the functionality described hereinabove with reference to Fig. 6A, the following cluster is selected for correction:

Exarts

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

Experts; Exerts; Exits

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'Exarts are'; 'Exarts are restoring'; 'Exarts are restoring the; 'Exarts are restoring the British'

Using the functionality described hereinabove with reference to Stage IIA, the matrix of frequencies of occurrence in an internet corpus seen in Table 26 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 26

CFS/ ALTERNATIVE CLUSTER CORRECTIONS	'Exarts are'	'Exarts are restoring'	'Exarts are restoring the'	'Exarts are restoring the British'
Experts	62540	0	0	0
Exerts	140	0	0	0
Exists	8225	0	0	0

All CFSs for which all alternative corrections have a zero frequency of occurrence are eliminated. In this example the following feature-grams are eliminated:

'Exarts are restoring'; 'Exarts are restoring the'; 'Exarts are restoring the British'

In this example the only remaining CFS is the feature-gram:

'Exarts are'

As seen from the foregoing example, the only CFS that survives the filtering process is "exarts are". As a result, the confidence level is relatively low, since the selection is based on only a single CFS, which is relatively short and includes, aside from the suspected word, only one word, which is a frequently occurring word.

Reference is now made to Fig. 9, which is a simplified flow chart illustrating functionality for context-based and word similarity-based scoring of various alternative corrections useful in the misused word and grammar correction functionality of Figs. 3, 10 and 11, and also in the vocabulary enhancement functionality of Fig. 4.

As seen in Fig. 9, the context-based and word similarity-based scoring of various alternative corrections proceeds in the following general stages:

I. NON-CONTEXTUAL SCORING - Various cluster alternatives are scored on the basis of similarity to a cluster in the input text in terms of their written appearance and sound similarity. This scoring does not take into account any contextual similarity outside of the given cluster.

II. CONTEXTUAL SCORING USING INTERNET CORPUS - Each of the various cluster alternatives is also scored on the basis of extracted contextual-feature-sequences (CFSs), which are provided as described hereinabove with reference to Fig. 5. This scoring includes the following sub-stages:

IIA. Frequency of occurrence analysis is carried out, preferably using an internet corpus, on the various alternative cluster corrections produced by the functionality of Figs. 7A or 7B, in the context of the CFSs extracted as described hereinabove in Fig. 5.

IIB. CFS selection and weighting of the various CFSs based on, inter alia, the results of the frequency of occurrence analysis of sub-stage IIA. Weighting is also based on relative inherent importance of various CFSs. It is appreciated that some of the CFSs may be given a weighting of zero and are thus not selected. The selected CFSs preferably are given relative weightings.

IIC. A frequency of occurrence metric is assigned to each alternative correction for each of the selected CFSs in sub-stage IIB.

IID. A reduced set of alternative cluster corrections is generated, based, inter alia, on the results of the frequency of occurrence analysis of sub-stage IIA, the frequency of occurrence metric of sub-stage IIC and the CFS selection and weighting of sub-stage IIB.

IIE. The input cluster is selected for use as a reference cluster correction.

IIF. A frequency of occurrence metric is assigned to the reference cluster correction of sub-stage IIE for each of the selected CFSs in stage IIB.

IIG. A ratio metric is assigned to each of the selected features in sub-stage IIB which represents the ratio of the frequency of occurrence metric for each alternative correction for that feature to the frequency of occurrence metric assigned to the reference cluster of sub-stage IIB.

III A most preferred alternative cluster correction is selected based on the results of stage I and the results of stage II.

IV. A confidence level score is assigned to the most preferred alternative cluster correction.

A more detailed description of the functionality described hereinabove in stages II-IV is presented hereinbelow:

With reference to sub-stage IIA, all of the CFSs which include the cluster to be corrected are generated as described hereinabove in Fig. 5. CFSs containing suspected errors, other than errors in the input cluster, are eliminated.

A matrix is generated indicating the frequency of occurrence in a corpus, preferably an internet corpus, of each of the alternative corrections for the cluster in each of the CFSs. All CFSs for which all alternative corrections have a zero frequency of occurrence are eliminated. Thereafter, all CFSs which are entirely included in other CFSs having at least a minimum threshold frequency of occurrence are eliminated.

The following example illustrates generation of a frequency of occurrence matrix:

The following input text is provided:

I lick two play outside a lot

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

lick two

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

like to; like two; lick two; lack two; lack true; like true

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'lick two'; 'I lick two'; 'lick two play'; 'I lick two play'; 'lick two play outside'; 'I lick two play outside'; 'lick two play outside a'

Using the functionality described hereinabove with reference to sub-stage IIA, the matrix of frequencies of occurrence in an internet corpus seen in Table 27 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 27

CFS/ ALTERNATIVE CLUSTER CORRECTIONS	lick two	I lick two	lick two play	I lick two play	lick two play outside	I lick two play outside	lick two play outside a
like to	75094300	3097250	432200	80500	1000	330	0
like two	327500	2400	0	0	0	0	0
lick two	600	0	0	0	0	0	0
lack two	1800	100	0	0	0	0	0
lack true	3400	90	0	0	0	0	0
like true	27500	350	0	0	0	0	0

All CFSs for which all alternative corrections have a zero frequency of occurrence are eliminated. In this example the following feature-gram is eliminated:

'lick two play outside a'

Thereafter, all CFSs which are entirely included in other CFSs having at least a minimum threshold frequency of occurrence are eliminated; For example the following feature-grams are eliminated:

'lick two'; 'I lick two'; 'lick two play'; 'I lick two play'; 'lick two play outside'

In this example the only remaining CFS is the following feature-gram:

'I lick two play outside'.

The resulting matrix appears as seen in Table 28:

TABLE 28

CFS/ALTERNATIVE CLUSTER CORRECTIONS	'I lick two play outside'
like to	330
like two	0
lick two	0
lack two	0
lack true	0
like true	0

The foregoing example illustrates the generation of a matrix in accordance with a preferred embodiment of the present invention. In this example, it is clear that "like to" is the preferred alternative correction. It is to be appreciated that in reality, the choices are not usually so straightforward. According, in further examples presented below, functionality is provided for making much more difficult choices among alternative corrections.

Returning to a consideration of sub-stage IIB, optionally each of the remaining CFSs is given a score as described hereinabove with reference to Fig. 5. Additionally CFSs which contain words introduced in an earlier correction iteration of the multi-word input and have a confidence level below a predetermined confidence level threshold are negatively biased.

In the general case, similarly to that described hereinabove in sub-stage IIC, preferably, a normalized frequency matrix is generated indicating the normalized frequency of occurrence of each CFS in the internet corpus. The normalized frequency matrix is normally generated from the frequency matrix by dividing each CFS frequency by a function of the frequencies of occurrence of the relevant cluster alternatives.

The normalization is operative to neutralize the effect of substantial differences in overall popularity of various alternative corrections. A suitable normalization factor is based on the overall frequencies of occurrence of various alternative corrections in a corpus as a whole, without regard to CFSs.

The following example illustrates the generation of a normalized frequency of occurrence matrix:

The following input text is provided typically by speech recognition:

footprints of a mysterious [hound/hand] said to be six feet tall

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

hound

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

hound; hand; sound

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'a mysterious hound'; 'hound said'

Using the functionality described hereinabove with reference to sub-stage IIC herein, the matrix of frequencies of occurrence and normalized frequencies of occurrence in an internet corpus seen in Table 29 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 29

	FREQUENCY			NORMALIZED FREQUENCY	
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**ALTERNATIVE CLUSTER CORRECTION**	**'a mysterious hound'**	**'hound said'**	**'a mysterious hound'**	**'hound said'**
Hound	542300	50	180	0.000092	0.000331
Hand	95073600	700	5900	0.000007	0.000062
Sound	59026500	590	600	0.000009	0.000010

It may be appreciated from the foregoing example that words having the highest frequencies of occurrence may not necessarily have the highest normalized frequencies of occurrence, due to substantial differences in overall popularity of various alternative corrections. In the foregoing example, "hound" has the highest normalized frequencies of occurrence and it is clear from the context of the input text that "hound" is the correct word, rather than "hand" which has higher frequencies of occurrence in the internet corpus.

It is a particular feature of the present invention that normalized frequencies, which neutralize substantial differences in overall popularity of various alternative corrections, are used in selecting among the alternative corrections. It is appreciated that other metrics of frequency of occurrence, other than normalized frequencies of occurrence, may alternatively or additionally be employed as metrics. Where the frequencies of occurrence are relatively low or particularly high, additional or alternative metrics are beneficial.

It will be appreciated from the discussion that follows that additional functionalities are often useful in selecting among various alternative corrections. These functionalities are described hereinbelow.

In sub-stage IID, each alternative cluster correction which is less preferred than another alternative correction according to both of the following metrics is eliminated:

i. having a word similarity score lower than the other alternative cluster correction; and

ii. having lower frequencies of occurrences and preferably also lower normalized frequencies of occurrence for all of the CFSs than the other alternative cluster correction.

The following example illustrates the elimination of alternative corrections as described hereinabove:

The following input text is provided:

I leave on a big house

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

leave on

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

leave in; live in; love in; leave on

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'I leave on a'; 'leave on a big'

Using the functionality described hereinabove with reference to Stage IIE herein, the matrix of frequencies of occurrence and normalized frequencies of occurrence in an internet corpus seen in Table 30 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 30

	FREQUENCY			NORMALIZED FREQUENCY	
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**ALTERNATIVE CLUSTER CORRECTION**	**'I leave on a'**	**'leave on a big'**	**'I leave on a'**	**'leave on a big'**
leave in	442650	1700	100	0.00384	0.00022
live in	98935570	267000	17800	0.00269	0.00018
love in	1023100	1880	320	0.00183	0.00031
leave on	267340	1040	0	0.00389	0

In this example, the non-contextual similarity scores of the alternative cluster corrections are as indicated in Table 31:

TABLE 31

ALTERNATIVE CLUSTER CORRECTION	**SIMILARITY SCORE**
leave in	0.90
live in	0.78
love in	0.67
leave on	1.00

The alternative cluster correction "love in" is eliminated as it has a lower similarity score as well as lower frequencies of occurrence and lower normalized frequencies of occurrence than "live in". The alternative cluster correction "leave in" is not eliminated at this stage since its similarity score is higher than that of "live in".

As can be appreciated from the foregoing, the result of operation of the functionality of sub-stage IID is a reduced frequency matrix and preferably also a reduced normalized frequency matrix, indicating the frequency of occurrence and preferably also the normalized frequency of occurrence of each of a reduced plurality of alternative corrections, each of which has a similarity score, for each of a reduced plurality of CFSs. The reduced set of alternative cluster corrections is preferably employed for all further alternative cluster selection functionalities as is seen from the examples which follow hereinbelow.

For each alternative correction in the reduced frequency matrix and preferably also in the reduced normalized frequency matrix, a final preference metric is generated. One or more of the following alternative metrics may be employed to generate a final preference score for each alternative correction:

The term "frequency function" is used below to refer to the frequency, the normalized frequency or a function of both the frequency and the normalized frequency.

A. One possible preference metric is the highest occurrence frequency function for each alternative cluster correction in the reduced matrix or matrices for any of the CFSs in the reduced matrix or matrices. For example, the various alternative cluster corrections would be scored as follows:

The following input text is provided:

I am vary satisfied with your work

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

vary

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

vary; very

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'am vary'; 'vary satisfied'; 'I am vary satisfied with'

Using the functionality described hereinabove with reference to sub-stage IIC herein, the matrix of frequencies of occurrence and normalized frequencies of occurrence in an internet corpus seen in Tables 32 and 33 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 32

	FREQUENCY			
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**ALTERNATIVE CLUSTER CORRECTION**	**'am vary'**	**'vary satisfied'**	**'I am vary satisfied with'**
Vary	20247200	800	70	0
Very	292898000	3123500	422700	30750

TABLE 33

	NORMALIZED FREQUENCY		
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**'am vary'**	**'vary satisfied'**	**'I am vary satisfied with'**
Vary	0.000039	0.000003	0
Very	0.010664	0.001443	0.000105

It is seen that in this example both from frequency of occurrence and normalized frequency of occurrence, "very" has the highest occurrence frequency function.

B. Another possible preference metric is the average occurrence frequency function of all CFSs for each alternative correction. For example, the various alternative corrections would be scored as follows:

The following input text is provided:

A while ago the lived 3 dwarfs

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

the

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

the; they; she; there

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'ago the lived'; 'the lived 3'

Using the functionality described hereinabove with reference to sub-stage IIC herein, the matrix of frequencies of occurrence, normalized frequencies of occurrence and average frequency of occurrence in an internet corpus seen in Tables 34 and 35 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 34

	FREQUENCY		
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**ALTERNATIVE CLUSTER CORRECTIONS**	**'ago the lived'**	**'the lived 3'**
The	19401194700	0	0
They	702221530	300	45
She	234969160	215	65
there	478280320	3200	40

TABLE 35

	NORMALIZED FREQUENCY		AVERAGE
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**'ago the lived'**	**'the lived 3'**	**Average frequency of occurrence**
The	0	0	0
They	0.0000004	0.00000006	172
She	0.0000009	0.00000027	140
there	0.0000066	0.00000008	1620

It is noted that "they" is selected based on the average frequency of occurrence, notwithstanding that "there" has a CFS whose frequency of occurrence is the maximum frequency of occurrence in the matrix.

In this example, the non-contextual similarity scores of the alternative cluster corrections are as indicated in Table 36:

TABLE 36

ALTERNATIVE CLUSTER CORRECTION	SIMILARITY SCORE
the	1.00
they	0.86
she	0.76
there	0.67

It is noted that the alternative cluster correction having the highest similarity score is not selected.

C. A further possible preference metric is the weighted sum over all CFSs for each alternative correction of the occurrence frequency function for each CFS multiplied by the score of that CFS as computed by the functionality described hereinabove with reference to Fig. 5.

D. A Specific Alternative Correction/CFS preference metric is generated, as described hereinabove with reference to sub-stages IIE - IIG, by any one or more, and more preferably most and most preferably all of the following operations on the alternative corrections in the reduced matrix or matrices:

i. The cluster from the original input text that is selected for correction is selected to be the reference cluster.

ii. A modified matrix is produced wherein in each preference matrix, the occurrence frequency function of each alternative correction in each feature gram is replaced by the ratio of the occurrence frequency function of each alternative correction to the occurrence frequency function of the reference cluster.

iii. A modified matrix of the type described hereinabove in ii. is further modified to replace the ratio in each preference metric by a function of the ratio which function reduces the computational importance of very large differences in ratios. A suitable such function is a logarithmic function. The purpose of this operation is to de-emphasize the importance of large differences in frequencies of occurrence in the final preference scoring of the most preferred alternative corrections, while maintaining the importance of large differences in frequencies of occurrence in the final preference scoring, and thus elimination, of the least preferred alternative corrections.

iv. A modified matrix of the type described hereinabove in ii or iii is additionally modified by multiplying the applicable ratio or function of ratio in each preference metric by the appropriate CFS score. This provides emphasis based on correct grammatical usage and other factors which are reflected in the CFS score.

v. A modified matrix of the type described hereinabove in ii, iii or iv is additionally modified by multiplying the applicable ratio or function of ratio in each preference metric by a function of a user uncertainty metric. Some examples of a user input uncertainty metric include the number of edit actions related to an input word or cluster performed in a word processor, vis-à-vis edit actions on other words of the

document; the timing of writing of an input word or cluster performed in a word processor, vis-à-vis time of writing of other words of the document and the timing of speaking of an input word or cluster performed in a speech recognition input functionality, vis-à-vis time of speaking of other words by this user. The user input uncertainty metric provides an indication of how certain the user was of this choice of words. This step takes the computed bias to a reference cluster and modifies it by a function of the user's certainty or uncertainty regarding this cluster.

vi. A modified matrix of the type described hereinabove in ii, iii, iv or v is additionally modified by generating a function of the applicable ratio, function of ratio, frequency of occurrence and normalized frequency of occurrence. A preferred function is generated by multiplying the applicable ratio or function of ratio in each preference metric by the frequency of occurrence of that CFS.

E. A final preference metric is computed for each alternative correction based on the Specific Alternative Correction/CFS preference metric as described hereinabove in D by multiplying the similarity score of the alternative correction by the sum of the Specific Alternative Correction/CFS preference metrics for all CFS for that Alternative Correction.

An example of such modified matrix is as follows:

The following input text is provided:

I will be able to teach base with you next week

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

teach

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

teach; touch

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'able to teach'; 'to teach base'

Using the functionality described hereinabove with reference to sub-stages IIA & IIC hereinabove, the matrix of frequencies of occurrence and normalized

frequencies of occurrence in an internet corpus seen in Table 37 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 37

	FREQUENCY			NORMALIZED FREQUENCY	
CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**ALTERNATIVE CLUSTER CORRECTION**	**'able to teach'**	**'to teach base'**	**'able to teach'**	**'to teach base'**
Teach	15124750	103600	40	0.00684	0.000002
touch	23506900	45050	27150	0.00191	0.001154

It is noted that for one feature, both the frequency of occurrence and the normalized frequency of occurrence of "teach" are greater than those of "touch", but for another feature, both the frequency of occurrence and the normalized frequency of occurrence of "touch" are greater than those of "teach". In order to make a correct choice of an alternative correction, ratio metrics, described hereinabove with reference to sub-stage IIG, are preferably employed as described hereinbelow.

In this example, the non-contextual similarity scores of the alternative cluster corrections are as indicated in Table 38:

TABLE 38

ALTERNATIVE CLUSTER CORRECTION	**SIMILARITY SCORE**
Teach	1.00
touch	0.89

It is seen that the reference cluster is "teach", since it has the highest similarity score. Nevertheless "touch" is selected based on the final preference score

described hereinabove. This is not intuitive as may be appreciated from a consideration of the above matrices which indicate that "teach" has the highest frequency of occurrence and the highest normalized frequency of occurrence. In this example, the final preference score indicates a selection of "touch" over "teach" since the ratio of frequencies of occurrence for a feature in which "touch" is favored is much greater than the ratio of frequencies of occurrence for the other feature in which "teach" is favored.

F. Optionally, an alternative correction may be filtered out on the basis of a comparison of frequency function values and preference metrics for that alternative correction and for the reference cluster using one or more of the following decision rules:

1. filtering out an alternative correction having a similarity score below a predetermined threshold and having a CFS frequency function that is less than the CFS frequency function of the reference cluster for at least one feature which has a CFS score which is higher than a predetermined threshold.

2. filtering out alternative corrections having a similarity score below a predetermined threshold and having a preference metric which is less than a predetermined threshold for at least one feature which has a CFS score which is higher than another predetermined threshold.

3. a. ascertaining the CFS score of each CFS;

b. for each CFS, ascertaining the CFS frequency functions for the reference cluster and for an alternative correction, thereby to ascertain whether the reference cluster or the alternative correction has a higher frequency function for that CFS;

c. summing the CFS scores of CFSs for which the alternative correction has a higher frequency than the reference cluster;

d. summing the CFS scores of CFSs for which the reference cluster has a higher frequency than the alternative correction;

e. if the sum in c. is less than the sum in d. filtering out that alternative correction.

The following example illustrates the filtering functionality described above.

The following input text is provided, typically by speech recognition functionality:

I want [two/to/too] items, please.

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

[two/to/too]

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

too; to; two

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'I want two'; 'want two items'

Using the functionality described hereinabove with reference to Stage IIA herein, the matrix of frequencies of occurrence in an internet corpus seen in Table 39 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 39

CFS/ALTERNATIVE CLUSTER CORRECTIONS	**'I want two'**	**'want two items'**
Too	9900	0
To	18286300	0
two	8450	140

The alternative corrections "too" and "to" are filtered out because they have zero frequency of occurrence for one of the CFSs, notwithstanding that they have high frequencies of occurrence of another CFS. Thus here, the only surviving CFS is "two".

G. As discussed hereinabove with reference to Stage III, a ranking is established based on the final preference metric developed as described hereinabove at

A - E on the alternative corrections which survive the filtering in F. The alternative correction having the highest final preference score is selected.

H. As discussed hereinabove with reference to Stage IV, a confidence level is assigned to the selected alternative correction. This confidence level is calculated based on one or more of the following parameters:

a. number, type and scoring of selected CFSs as provided in sub-stage IIB above;

b. statistical significance of frequency of occurrence of the various alternative cluster corrections, in the context of the CFSs;

c. degree of consensus on the selection of an alternative correction, based on preference metrics of each of the CFSs and the word similarity scores of the various alternative corrections;

d. non-contextual similarity score (stage I) of the selected alternative cluster correction being above a predetermined minimum threshold.

e. extent of contextual data available, as indicated by the number of CFSs in the reduced matrix having CFS scores above a predetermined minimum threshold and having preference scores over another predetermined threshold.

If the confidence level is above a predetermined threshold, the selected alternative correction is implemented without user interaction. If the confidence level is below the predetermined threshold but above a lower predetermined threshold, the selected alternative correction is implemented but user interaction is invited. If the confidence level is below the lower predetermined threshold, user selection based on a prioritized list of alternative corrections is invited.

The following examples are illustrative of the use of confidence level scoring:

The following input text is provided:

He was not feeling wale when he returned

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

wale

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

wale; well

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'was not feeling wale'; 'not feeling wale when'; 'feeling wale when he'; 'wale when he returned'

Using the functionality described hereinabove with reference to sub-stage IIA herein, the matrix of frequencies of occurrence in an internet corpus seen in Table 40 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 40

CFS/ALTERNATIVE CLUSTER CORRECTIONS	**'was not feeling wale'**	**'not feeling wale when'**	**'feeling wale when he'**	**'wale when he returned'**
Wale	0	0	0	0
Well	31500	520	100	140

The foregoing example illustrates that, according to all the criteria set forth in H above, the selection of 'well' over 'wale' has a high confidence level.

In the following example, the confidence level is somewhat less, due to the fact that the alternative correction 'back' has a higher frequency of occurrence than 'beach' in the CFS 'beech in the summer' but 'beach' has a higher frequency of occurrence than 'back' in the CFSs 'on the beech in' and 'the beech in the'. The alternative correction 'beach' is selected with an intermediate confidence level based on criterion H(c).

The following input text is provided:

I like to work on the beech in the summer

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

beech

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

beach; beech; back

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'on the beech in'; 'the beech in the'; 'beech in the summer'

Using the functionality described hereinabove with reference to Stage IIA, the matrix of frequencies of occurrence in an internet corpus seen in Table 41 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 41

CFS/ALTERNATIVE CLUSTER CORRECTIONS	**'on the beech in'**	**'the beech in the'**	**'beech in the summer'**
Beach	110560	42970	2670
Beech	50	55	0
Back	15300	10390	20090

The alternative correction 'beach' is selected with an intermediate confidence level based on criterion H(c).

In the following example, the confidence level is even less, based on criterion H(a):

The following input text is received:

Exerts are restoring the British Museum's round reading room

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

Exerts

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

Expert; Exerts; Exits

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'Exerts are'; 'Exerts are restoring'; 'Exerts are restoring the'; 'Exerts are restoring the British'

Using the functionality described hereinabove with reference to sub-stage IIA, the matrix of frequencies of occurrence in an internet corpus seen in Table 42 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 42

CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**'Exerts are'**	**'Exerts are restoring'**	**'Exerts are restoring the'**	**'Exerts are restoring the British'**
Experts	62540	0	0	0
Exerts	140	0	0	0
Exists	8225	0	0	0

All CFSs for which all alternative corrections have a zero frequency of occurrence are eliminated. In this example the following feature-grams are eliminated:

'Exerts are restoring'; 'Exerts are restoring the'; 'Exerts are restoring the British'

In this example the only remaining CFS is the feature-gram:

'Exerts are'

As seen from the foregoing example, the only CFS that survives the filtering process is 'Exerts are'. As a result, the confidence level is relatively low, since the selection is based on only a single CFS, which is relatively short and includes, aside from the suspected word, only one word, which is a frequently occurring word.

The following example illustrates the usage of the final preference score metric described in stages D & E above.

The following input text is provided:

Some kids don't do any sport and sit around doing nothing and getting fast so you will burn some calories and get a lot fitter if you exercise.

Using the functionality described hereinabove with reference to Fig. 6B, the following cluster is selected for correction:

fast

Using the functionality described hereinabove with reference to Fig. 7A, the following alternative cluster corrections are generated (partial list):

fat; fast

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'and getting fast'; 'getting fast so'; 'fast so you'; 'fast so you will'

Using the functionality described hereinabove with reference to sub-stage IIA herein, the matrix of frequencies of occurrence in an internet corpus seen in Table 43 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 43

CFS/ALTERNATIVE CLUSTER CORRECTIONS	**'and getting fast'**	**'getting fast so'**	**'fast so you'**	**'fast so you will'**
CFS IMPORTANCE SCORE	0.8	0.8	0.05	0.2
Fast	280	20	6500	250
Fat	1960	100	1070	115

In this example, the non-contextual similarity scores of the alternative cluster corrections are as indicated in Table 44:

TABLE 44

ALTERNATIVE CLUSTER CORRECTION	SIMILARITY SCORE
fast	1
fat	0.89

Using the final preference score metric described in stages D & E above, the alternative correction "fat" is selected with low confidence.

Reference is now made to Fig. 10, which is a detailed flowchart illustrating the operation of missing item correction functionality. The missing item correction functionality is operative to correct for missing articles, prepositions, punctuation and other items having principally grammatical functions in an input text. This functionality preferably operates on a spelling-corrected input text output from the spelling correction functionality of Fig. 1.

Identification of suspected missing items is carried out preferably in the following manner:

Initially, feature-grams are generated for a spelling-corrected input text. The frequency of occurrence of each feature-gram in the spelling-corrected input text in a corpus, preferably an internet corpus (FREQ F-G), is ascertained.

An expected frequency of occurrence of each feature-gram (EFREQ F-G) is calculated as follows:

A feature-gram is assumed to contain n words, identified as W_1 - W_n.

W_i designates the i'th word in the feature-gram

An expected frequency of occurrence of a given feature-gram is taken to be the highest of expected frequencies of that feature-gram based on division of the words in the feature-gram into two consecutive parts following each of the words W_1 ... $W_{(n-1)}$.

The expected frequency of a feature-gram based on division of the words in the feature-gram into two consecutive parts following a word W_i can be expressed as follows:

EFREQ F-G in respect of W_i = (FREQ (W_1 - W_i) * FREQ (W_{i+1} - W_n))/(TOTAL OF FREQUENCIES OF ALL WORDS IN THE CORPUS)

The expected frequencies of each feature-gram based on all possible divisions of the words in the feature-gram into two consecutive parts are calculated.

If FREQ F-G/EFREQ F-G in respect of W_i is less than a predetermined threshold, the feature-gram in respect of W_i is considered to be suspect in terms of there being a missing article, preposition or punctuation between W_i and W_{i+1} in that feature gram.

A suspect word junction between two consecutive words in a spelling-corrected input text is selected for correction, preferably by attempting to find the word junction which is surrounded by the largest amount of non-suspected contextual data. Preferably, the word junction that has the longest sequence or sequences of non-suspected word junctions in its vicinity is selected.

One or, preferably, more alternative insertions is generated for each word junction, preferably based on a predefined set of possibly missing punctuation, articles, prepositions, conjunctions or other items, which normally do not include nouns, verbs or adjectives.

At least partially context-based and word similarity-based scoring of the various alternative insertions is provided, preferably based on a correction alternatives scoring algorithm, described hereinabove with reference to Fig. 9 and hereinbelow.

The following example is illustrative:

The following input text is provided:

I can't read please help me

Using the functionality described hereinabove with reference to Fig. 5, the following feature-grams are generated (partial list):

I can't read; can't read please; read please help; please help me

Using the functionality described hereinabove, a matrix of the frequencies of occurrence in an internet corpus is generated for the above list of feature-grams which typically appears as seen in Table 45:

TABLE 45

FEATURE-GRAM	FREQUENCY OF OCCURRENCE
I can't read	5600
can't read please	0
read please help	55
please help me	441185

The expected frequency of occurrence is calculated for each feature-gram in respect of each word W_i in the feature-gram, in accordance with the following expression:

EFREQ F-G in respect of W_i = (FREQ (W_1 - W_i) * FREQ (W_{i+1} - W_n))/(TOTAL OF FREQUENCIES OF ALL WORDS IN THE CORPUS)

The exemplary results of some of these calculations are seen in Tables 46 and 47:

TABLE 46

FEATURE-GRAM	FREQUENCY OF OCCURRENCE	EXPECTED FREQUENCY OF OCCURRENCE WITH RESPECT TO "read"	FREQ F-G in respect of "read"/ EFREQ F-G in respect of "read"
can't read please	0	0	0
read please help	55	220	0.25

TABLE 47

FEATURE-GRAM	FREQUENCY OF OCCURRENCE
read	157996585
please help	1391300

As seen from the above results, the actual frequency of occurrence of each of the feature-grams is less than the expected frequency of occurrence thereof. This indicates suspected absence of an item, such as punctuation.

A list of alternative insertions to follow the word "read" is generated. This list preferably includes a predetermined list of punctuation, articles, conjunctions and prepositions. Specifically, it will include a period "."

A partial list of the alternatives is:

'read please'; 'read. Please'; 'read of please'; 'read a please'

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated:

'I can't read [?]'; 'read [?] please help'; ' [?] please help me'

Using the functionality described in stage IIA of Fig. 9, the matrix of frequencies of occurrence in an internet corpus seen in Table 48 is generated for the above list of alternative cluster corrections in the above list of CFSs:

When a '.' is included in a cluster, the CFS frequency of occurrence that includes the cluster with the '.' is retrieved separately for the text before and after the '.'. i.e., the feature-gram "can't read. Please" will not be generated because it includes two separate grammar parsing phrases.

TABLE 48

CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**'can't read [?]'**	**'read [?] please help'**	**'[?] please help me'**
read please	0	0	0
read. Please	1093	0	357945*
read of please	0	0	0
read a please	0	0	0

* Note: A '.' is omitted from the beginning of a feature gram when calculating its frequency of occurrence in the corpus. For example, the frequency of ". Please help me" is identical to the frequency of "Please help me".

Using the functionality described in stages D & E of Fig. 9 the final preference metric selects the alternative correction "read. Please" and the corrected input text is:

I can't read. Please help me.

The following example illustrates the functionality of adding a missing preposition.

The following input text is provided:

I sit the sofa

Using the functionality described hereinbelow, the following cluster is selected for correction:

'sit the'

Using the functionality described hereinbelow, the following alternative cluster corrections are generated (partial list):

sit on the; sit of the; sit the

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated:

'I sit the'; 'sit the sofa'

Using the functionality described in stage IIA with reference to Fig. 9, the matrix of frequencies of occurrence in an internet corpus seen in Table 49 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 49

CFS/ALTERNATIVE CLUSTER CORRECTIONS	**'I sit [?] the'**	**'sit [?] the sofa'**
sit on the	26370	7400
sit of the	0	0
sit the	2100	0

Using the functionality described in stages IID & IIE of Fig. 9 the final preference metric selects the alternative correction "sit on the" and the corrected input text is:

I sit on the sofa.

Reference is now made to Fig. 11, which is a detailed flowchart illustrating the operation of superfluous item correction functionality. The superfluous item correction functionality is operative to correct for superfluous articles, prepositions, punctuation and other items having principally grammatical functions in an input text. This functionality preferably operates on a spelling-corrected input text output from the spelling correction functionality of Fig. 1.

It is appreciated that the functionality of Fig. 11 may be combined with the functionality of Fig. 10 or alternatively carried out in parallel therewith, prior thereto or following operation thereof.

Identification of suspected superfluous items is carried out preferably in the following manner:

A search is carried out on the spelling-corrected input text to identify items belonging to a predefined set of possibly superfluous punctuation, articles, prepositions, conjunctions and other items, which normally do not include nouns, verbs or adjectives.

For each such item, feature-grams are generated for all portions of the misused-word and grammar corrected, spelling-corrected input text containing such item. A frequency of occurrence is calculated for each such feature-gram and for a corresponding feature-gram in which the item is omitted.

If the frequency of occurrence for the feature-gram in which the item is omitted exceeds the frequency of occurrence for the corresponding feature-gram in which the item is present, the item is considered as suspect.

A suspect item in a misused-word and grammar corrected, spelling-corrected input text is selected for correction, preferably by attempting to find the item which is surrounded by the largest amount of non-suspected contextual data. Preferably, the item that has the longest sequence or sequences of non-suspected words in its vicinity is selected.

A possible item deletion is generated for each suspect item. At least partially context-based and word similarity-based scoring of the various alternatives, i.e. deletion of the item or non-deletion of the item, is provided, preferably based on a correction alternatives scoring algorithm, described hereinabove with reference to Fig. 9 and hereinbelow.

The following example is illustrative.

The following input text is provided:

It is a nice, thing to wear.

The input text is searched to identify any items which belong to a predetermined list of commonly superfluous items, such as, for example, punctuation, prepositions, conjunctions and articles.

In this example, the comma "," is identified as belonging to such a list.

Using the functionality described hereinabove with reference to Fig. 5, the feature-grams, seen in Table 50, which include a comma "," are generated and identical feature-grams without the comma are also generated (partial list):

TABLE 50

FEATURE-GRAM WITH COMMA	FEATURE-GRAM WITHOUT COMMA
is a nice, thing	is a nice thing
a nice, thing to	a nice thing to
nice, thing to wear	nice thing to wear

Using the functionality described hereinabove, a matrix of the frequencies of occurrence in an internet corpus is generated for the above list of feature-grams which typically appears as seen in Table 51:

TABLE 51

FEATURE-GRAM WITH COMMA	FREQUENCY OF OCCURRENCE OF FEATURE-GRAM WITH COMMA	FEATURE-GRAM WITHOUT COMMA	FREQUENCY OF OCCURRENCE OF FEATURE-GRAM WITHOUT COMMA
is a nice, thing	0	is a nice thing	10900
a nice, thing to	0	a nice thing to	39165
nice, thing to wear	0	nice thing to wear	100

As seen in the matrix above, the frequency of occurrence for the feature grams with the "," omitted exceeds the frequency of occurrence for corresponding feature grams with the "," present. Therefore, the "," is considered as suspect of being superfluous.

The possible deletion of the comma is considered, based on context based scoring of the following alternatives of keeping the comma and omitting the comma:

'nice,' ; 'nice'

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'a nice,'; 'nice, thing'; 'is a nice,'; 'a nice, thing'; 'nice, thing to'

Using the functionality described hereinabove with reference to figure 9 Stage IIA, the matrix of frequencies of occurrence in an internet corpus seen in Table 52 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 52

CFS/ ALTERNATIVE CLUSTER CORRECTIONS	**'a nice,'**	**'nice, thing'**	**'is a nice,'**	**'a nice, thing'**	**'nice, thing to'**
nice,	379400	0	37790	0	0
Nice	11809290	300675	1127040	69100	58630

All CFSs which are entirely included in other CFSs having at least a minimum threshold frequency of occurrence are eliminated. For example the following feature-grams are eliminated:

'a nice,'; 'nice, thing'

In this example the remaining CFSs are the feature-grams:

'is a nice,'; 'a nice, thing'; 'nice, thing to'

Using the final preference score described in stages D & E of Fig. 9 above, the alternative correction "nice", without the comma, is selected. The input text after the comma deletion is:

It is a nice thing to wear.

The following example illustrates the functionality of removing a superfluous article.

The following input text is provided:

We should provide them a food and water.

Using the functionality described hereinabove with reference to Fig. 11, the following cluster is selected for correction:

a food

Using the functionality described hereinabove with reference to Fig. 11, the following alternative cluster corrections are generated (partial list):

a food; food

Using the functionality described hereinabove with reference to Fig. 5, the following CFSs are generated (partial list):

'provide them a food'; 'them a food and'; 'a food and water'

Using the functionality described hereinabove with reference to sub-stage IIA herein, the matrix of frequencies of occurrence in an internet corpus seen in Table 53 is generated for the above list of alternative cluster corrections in the above list of CFSs:

TABLE 53

CFS/ALTERNATIVE CLUSTER CORRECTIONS	**'provide them a food'**	**'them a food and'**	**'a food and water'**
a food	0	0	950
Food	790	12775	415620

Using the scoring functionality described in Fig. 9, the final preference metric selects the alternative correction "food" and the corrected input text is:

We should provide them food and water.

It will be appreciated by persons skilled in the art that the present invention is not limited to what has been particularly shown and described hereinabove. Rather the scope of the present invention includes both combinations and sub-combinations of the various features described and shown hereinabove and modifications thereof which will occur to persons skilled in the art upon reading the foregoing description and which are not in the prior art.

CLAIMS

1. A computer-assisted language correction system comprising:

an alternatives generator, generating on the basis of an input sentence a text-based representation providing multiple alternatives for each of a plurality of words in the sentence;

a selector for selecting among at least said multiple alternatives for each of said plurality of words in the sentence, based at least partly on an internet corpus; and

a correction generator operative to provide a correction output based on selections made by said selector.

2. A computer-assisted language correction system according to claim 1 wherein said selector is operative to make said selections based on at least one of the following correction functions:

spelling correction;

misused word correction;

grammar correction; and

vocabulary enhancement.

3. A computer-assisted language correction system according to claim 1 wherein said selector is operative to make said selections based on at least two of the following correction functions:

spelling correction;

misused word correction;

grammar correction; and

vocabulary enhancement.

4. A computer-assisted language correction system according to claim 3 and wherein said selector is operative to make said selections based on at least one of the following time ordering of corrections:

spelling correction prior to at least one of misused word correction, grammar correction and vocabulary enhancement; and

misused word correction and grammar correction prior to vocabulary enhancement.

5. A computer-assisted language correction system according to any of claims 2 - 4 and wherein:

said input sentence is provided by one of the following functionalities:

word processor functionality;

machine translation functionality;

speech-to-text conversion functionality;

optical character recognition functionality; and

instant messaging functionality; and

said selector is operative to make said selections based on at least one of the following correction functions:

misused word correction;

grammar correction; and

vocabulary enhancement.

6. A computer-assisted language correction system according to any of claims 2 - 5 and wherein said correction generator comprises a corrected language input generator operative to provide a corrected language output based on selections made by said selector without requiring user intervention.

7. A computer-assisted language correction system according to any of claims 2-6 and wherein said grammar correction functionality includes at least one of punctuation, verb inflection, single/plural, article and preposition correction functionalities.

8. A computer-assisted language correction system according to any of claims 2-7 and wherein said grammar correction functionality includes at least one of replacement, insertion and omission correction functionalities.

9. A computer-assisted language correction system according to any of claims 1-8 and wherein said selector includes context based scoring functionality operative to rank said multiple alternatives, based at least partially on contextual feature-sequence (CFS) frequencies of occurrences in an internet corpus.

10. A computer-assisted language correction system according to claim 9 and wherein said context based scoring functionality is also operative to rank said multiple alternatives based at least partially on normalized CFS frequencies of occurrences in said internet corpus.

11. A computer-assisted language correction system comprising:

at least one of:

spelling correction functionality;

misused word correction functionality;

grammar correction functionality; and

vocabulary enhancement functionality; and

contextual feature-sequence functionality cooperating with at least one of said spelling correction functionality; said misused word correction functionality, grammar correction functionality and said vocabulary enhancement functionality and employing an internet corpus.

12. A computer-assisted language correction system according to claim 11 and wherein said grammar correction functionality includes at least one of punctuation, verb inflection, single/plural, article and preposition correction functionalities.

13. A computer-assisted language correction system according to claim 11 or claim 12 and wherein said grammar correction functionality includes at least one of replacement, insertion and omission correction functionalities.

14. A computer-assisted language correction system according to any of claims 11 - 13 and comprising:

at least two of:

said spelling correction functionality;

said misused word correction functionality;

said grammar correction functionality; and

said vocabulary enhancement functionality, and

wherein said contextual feature-sequence functionality cooperates with at least two of said spelling correction functionality; said misused word correction functionality, said grammar correction functionality and said vocabulary enhancement functionality and employing an internet corpus.

15. A computer-assisted language correction system according to any of claims 11 - 13 and comprising:

at least three of:

said spelling correction functionality;

said misused word correction functionality;

said grammar correction functionality;

said vocabulary enhancement functionality; and

wherein said contextual feature-sequence functionality cooperates with at least three of said spelling correction functionality; said misused word correction functionality, said grammar correction functionality and said vocabulary enhancement functionality and employing an internet corpus.

16. A computer-assisted language correction system according to any of claims 11 - 13 and comprising:

said spelling correction functionality;

said misused word correction functionality;

said grammar correction functionality; and

said vocabulary enhancement functionality; and

wherein said contextual feature-sequence functionality cooperates with said spelling correction functionality; said misused word correction functionality, said grammar correction functionality and said vocabulary enhancement functionality and employing an internet corpus.

17. A computer-assisted language correction system according to any of claims 11 - 16 and wherein said correction generator comprises a corrected language generator operative to provide a corrected language output based on selections made by said selector without requiring user intervention.

18. A computer-assisted language correction system comprising:

an alternatives generator, generating on the basis of a language input a text-based representation providing multiple alternatives for each of a plurality of words in the sentence;

a selector for selecting among at least said multiple alternatives for each of said plurality of words in the language input, based at least partly on a relationship between selected ones of said multiple alternatives for at least some of said plurality of words in said language input; and

a correction generator operative to provide a correction output based on selections made by said selector.

19. A computer-assisted language correction system according to claim 18 and wherein said language input comprises at least one of an input sentence and an input text.

20. A computer-assisted language correction system according to claim 18 or claim 19 and wherein said language input is speech and said generator converts said

language input in speech to a text-based representation providing multiple alternatives for a plurality of words in said language input.

21. A computer-assisted language correction system according to claim 18 or claim 19 and wherein:

said language input is at least one of:

a text input;

an output of optical character recognition functionality;

an output of machine translation functionality; and

an output of word processing functionality; and

said generator converts said language input in text to a text-based representation providing multiple alternatives for a plurality of words in the language input.

22. A computer-assisted language correction system according to any of claims 18 - 21 and wherein said selector is operative to make said selections based on at least two of the following correction functions:

spelling correction;

misused word correction;

grammar correction; and

vocabulary enhancement.

23. A computer-assisted language correction system according to claim 22 and wherein said selector is operative to make said selections based on at least one of the following time ordering of corrections:

spelling correction prior to at least one of misused word correction, grammar correction and vocabulary enhancement; and

misused word correction and grammar correction prior to vocabulary enhancement.

24. A computer-assisted language correction system according to any of claims 18-20 and wherein said language input is speech and said selector is operative to make said selections based on at least one of the following correction functions:

misused word correction;

grammar correction; and

vocabulary enhancement.

25. A computer-assisted language correction system according to any of claims 18 - 24 and wherein said selector is operative to make said selections by carrying out at least two of the following functions:

selection of a first set of words or combinations of words which include less than all of said plurality of words in said language input for an initial selection;

thereafter ordering elements of said first set of words or combinations of words to establish priority of selection; and

thereafter when selecting among said multiple alternatives for an element of said first set of words, choosing other words, but not all, of said plurality of words as a context to influence said selecting.

26. A computer-assisted language correction system according to any of claims 18 - 25 and wherein said selector is operative to make said selections by carrying out the following function:

when selecting for an element having at least two words, evaluating each of said multiple alternatives for each of said at least two words in combination with each of said multiple alternatives for each other of said at least two words.

27. A computer-assisted language correction system according to any of claims 18 - 26 and wherein said correction generator comprises a corrected language input generator operative to provide a corrected language output based on selections made by said selector without requiring user intervention.

28. A computer-assisted language correction system comprising:

a misused-word suspector evaluating at least most of the words in an language input on the basis of their fit within a context of the language input; and

a correction generator operative to provide a correction output based at least partially on an evaluation performed by said suspector.

29. A computer-assisted language correction system according to claim 28 and also comprising:

an alternatives generator, generating on the basis of said language input, a text-based representation providing multiple alternatives for at least one of said at least most words in said language input; and

a selector for selecting among at least said multiple alternatives for each of said at least one of said at least most words in said language input, and wherein

said correction generator is operative to provide said correction output based on selections made by said selector.

30. A computer-assisted language correction system according to claim 28 or 29 and also comprising:

a suspect word output indicator indicating an extent to which at least some of said at least most of said words in said language input is suspect as a misused-word.

31. A computer-assisted language correction system according to any of claims 28 - 30 and wherein said correction generator comprises an automatic corrected language generator operative to provide a corrected text output based at least partially on an evaluation performed by said suspector, without requiring user intervention.

32. A computer-assisted language correction system according to any of claims 28 - 31 and wherein said language input is speech and said selector is operative to make said selections based on at least one of the following correction functions:

misused word correction;

grammar correction; and

vocabulary enhancement.

33. A computer-assisted language correction system comprising:

a misused-word suspector evaluating words in an language input;

an alternatives generator, generating multiple alternatives for at least some of the words in the language input evaluated as suspect words by said suspector, at least one of said multiple alternatives for a word in the language input being consistent with a contextual feature of said word in the language input in an internet corpus;

a selector for selecting among at least said multiple alternatives; and

a correction generator operative to provide a correction output based at least partially on a selection made by said selector.

34. A computer-assisted language correction system comprising:

a misused-word suspector evaluating words in an language input and identifying suspect words;

an alternatives generator, generating multiple alternatives for said suspect words;

a selector, grading each said suspect word as well as ones of said multiple alternatives therefor generated by said alternatives generator according to multiple selection criteria, and applying a bias in favor of said suspect word vis-à-vis ones of said multiple alternatives therefor generated by said alternatives generator; and

a correction generator operative to provide a correction output based at least partially on a selection made by said selector.

35. A computer-assisted language correction system comprising:

an alternatives generator, generating on the basis of an input multiple alternatives for at least one word in the input;

a selector, grading each said at least one word as well as ones of said multiple alternatives therefor generated by said alternatives generator according to multiple selection criteria, and applying a bias in favor of said at least one word vis-à-vis ones of said multiple alternatives therefor generated by said alternatives generator,

said bias being a function of an input uncertainty metric indicating uncertainty of a person providing said input; and

a correction generator operative to provide a correction output based on a selection made by said selector.

36. A computer-assisted language correction system comprising:

an incorrect word suspector evaluating at least most of the words in a language input, said suspector being at least partially responsive to an input uncertainty metric indicating uncertainty of a person providing said input, said suspector providing a suspected incorrect word output; and

an alternatives generator, generating a plurality of alternatives for suspected incorrect words identified by said suspected incorrect word output;

a selector for selecting among each suspected incorrect word and the plurality of alternatives generated by said alternatives generator; and

a correction generator operative to provide a correction output based on a selection made by said selector.

37. A computer-assisted language correction system comprising:

at least one of a spelling correction module, a misused-word correction module, a grammar correction module and a vocabulary enhancement module receiving a multi-word input and providing a correction output, each of said at least one of a spelling correction module, a misused-word correction module, a grammar correction module and a vocabulary enhancement module including:

an alternative word candidate generator including:

phonetic similarity functionality operative to propose alternative words based on phonetic similarity to a word in said input and to indicate a metric of phonetic similarity; and

character string similarity functionality operative to propose alternative words based on character string similarity to a word in said input and to indicate a metric of character string similarity for each alternative word; and

a selector operative to select either a word in said output or an alternative word candidate proposed by said alternative word candidate generator by employing said phonetic similarity and character string similarity metrics together with context-based selection functionality.

38. A computer-assisted language correction system comprising:

suspect word identification functionality, receiving a multi-word language input and providing a suspect word output which indicates suspect words;

feature identification functionality operative to identify features including said suspect words;

an alternative selector identifying alternatives to said suspect words;

feature occurrence functionality employing a corpus and providing an occurrence output, ranking various features including said alternatives as to their frequency of use in the corpus; and

a selector employing said occurrence output to provide a correction output,

said feature identification functionality comprising feature filtration functionality including at least one of:

functionality for eliminating features containing suspected errors;

functionality for negatively biasing features which contain words introduced in an earlier correction iteration of said multi-word input and which have a confidence level below a confidence level predetermined threshold; and

functionality for eliminating features which are contained in another feature having an frequency of occurrence above a predetermined frequency threshold.

39. A computer-assisted language correction system according to any of claims 33-38 and wherein said selector is operative to make said selections based on at least two of the following correction functions:

spelling correction;

misused word correction;

grammar correction; and

vocabulary enhancement.

40. A computer-assisted language correction system according to claim 39 and wherein said selector is operative to make said selections based on at least one of the following time ordering of corrections:

spelling correction prior to at least one of misused word correction, grammar correction and vocabulary enhancement; and

misused word correction and grammar correction prior to vocabulary enhancement.

41. A computer-assisted language correction system according to any of claims 33 - 40 and wherein said language input is speech and said selector is operative to make said selections based on at least one of the following correction functions:

grammar correction; and

misused word correction; and

vocabulary enhancement.

42. A computer-assisted language correction system according to any of claims 33 - 41 and wherein said correction generator comprises a corrected language input generator operative to provide a corrected language output based on selections made by said selector without requiring user intervention.

43. A computer-assisted language correction system according to any of claims 33 – 34 and 37 – 42 and wherein said selector is also operative to make said selections based at least partly on a user input uncertainty metric.

44. A computer-assisted language correction system according to claim 43 and wherein said user input uncertainty metric is a function based on a measurement of the uncertainty of a person providing said input.

45. A computer-assisted language correction system according to any of claims 33 - 44 and wherein said selector also employs user input history learning functionality.

46. A computer-assisted language correction system comprising:

suspect word identification functionality, receiving a multi-word language input and providing a suspect word output which indicates suspect words;

feature identification functionality operative to identify features including said suspect words;

an alternative selector identifying alternatives to said suspect words;

occurrence functionality employing a corpus and providing an occurrence output, ranking features including said alternatives as to their frequency of use in the corpus; and

a correction output generator, employing said occurrence output to provide a correction output,

said feature identification functionality comprising:

at least one of:

N-gram identification functionality; and

co-occurrence identification functionality; and

at least one of:

skip-gram identification functionality;

switch-gram identification functionality; and

previously used by user feature identification functionality.

47. A computer-assisted language correction system comprising:

a grammatical error suspector evaluating at least most of the words in an language input on the basis of their fit within a context of the language input; and

a correction generator operative to provide a correction output based at least partially on an evaluation performed by said suspector.

48. A computer-assisted language correction system according to claim 47 and also comprising:

an alternatives generator, generating on the basis of said language input, a text-based representation providing multiple alternatives for at least one of said at least most words in said language input; and

a selector for selecting among at least said multiple alternatives for each of said at least one of said at least most words in said language input, and wherein

said correction generator is operative to provide said correction output based on selections made by said selector.

49. A computer-assisted language correction system according to claim 47 or 48 and also comprising:

a suspect word output indicator indicating an extent to which at least some of said at least most of said words in said language input is suspect as containing grammatical error.

50. A computer-assisted language correction system according to any of claims 47 - 49 and wherein said correction generator comprises an automatic corrected language generator operative to provide a corrected text output based at least partially on an evaluation performed by said suspector, without requiring user intervention.

51. A computer-assisted language correction system comprising:

a grammatical error suspector evaluating words in a language input;

an alternatives generator, generating multiple alternatives for at least some of the words in the language input evaluated as suspect words by said suspector, at least one of said multiple alternatives for a word in the language input being consistent with a contextual feature of said word in the language input;

a selector for selecting among at least said multiple alternatives; and

a correction generator operative to provide a correction output based at least partially on a selection made by said selector.

52. A computer-assisted language correction system comprising:

a grammatical error suspector evaluating words in an language input and identifying suspect words;

an alternatives generator, generating multiple alternatives for said suspect words;

a selector, grading each said suspect word as well as ones of said multiple alternatives therefor generated by said alternatives generator according to multiple selection criteria, and applying a bias in favor of said suspect word vis-à-vis ones of said multiple alternatives therefor generated by said alternatives generator; and

a correction generator operative to provide a correction output based at least partially on a selection made by said selector.

53. A computer-assisted language correction system according to claim 51 or claim 52 and wherein said correction generator comprises a corrected language input generator operative to provide a corrected language output based on selections made by said selector without requiring user intervention.

54. A computer-assisted language correction system comprising context based scoring of various alternative corrections, based at least partially on contextual feature-sequence (CFS) frequencies of occurrences in an internet corpus.

55. A computer-assisted language correction system according to claim 54 and also comprising at least one of:

spelling correction functionality;

misused word correction functionality;

grammar correction functionality; and

vocabulary enhancement functionality,

cooperating with said context based scoring.

56. A computer-assisted language correction system according to claim 54 or claim 55 and wherein said context based scoring is also based at least partially on normalized CFS frequencies of occurrences in an internet corpus.

57. A computer-assisted language correction system according to any of claims 54 – 56 and wherein said context based scoring is also based at least partially on a CFS importance score.

58. A computer-assisted language correction system according to any of claims 54 - 57 and wherein said CFS importance score is a function of at least one of the following:

operation of a part-of-speech tagging and sentence parsing functionality; a CFS length; a frequency of occurrence of each of the words in the CFS and a CFS type.

59. A computer-assisted language correction system comprising vocabulary enhancement functionality including:

vocabulary-challenged words identification functionality;

alternative vocabulary enhancements generation functionality; and

context based scoring functionality, based at least partially on contextual feature-sequence (CFS) frequencies of occurrences in an internet corpus,

said alternative vocabulary enhancements generation functionality including thesaurus pre-processing functionality operative to generate alternative vocabulary enhancements.

60. A computer-assisted language correction system comprising:

an alternatives generator, generating on the basis of an input sentence a text-based representation providing multiple alternatives for each of a plurality of words in the sentence;

a selector for selecting among at least said multiple alternatives for each of said plurality of words in the sentence;

a confidence level assigner operative to assign a confidence level to a selected alternative from said multiple alternatives; and

a correction generator operative to provide a correction output based on selections made by said selector and at least partially on said confidence level.

61. A computer-assisted language correction system according to claim 60 and wherein said multiple alternatives are evaluated based on contextual feature sequences (CFSs) and said confidence level is based on at least one of the following parameters:

number, type and scoring of selected CFSs;

a measure of statistical significance of frequency of occurrence of said multiple alternatives, in the context of the CFSs;

degree of consensus on the selection of one of said multiple alternatives, based on preference metrics of each of said CFSs and word similarity scores of said multiple alternatives;

a non-contextual similarity score of said one of said multiple alternatives being above a first predetermined minimum threshold; and

an extent of contextual data available, as indicated by the number of said CFSs having CFS scores above a second predetermined minimum threshold and having preference scores over a third predetermined threshold.

62. A computer-assisted language correction system comprising:

a punctuation error suspector evaluating at least some of the words and punctuation in a language input on the basis of their fit within a context of the language input based on frequency of occurrence of feature-grams of said language input in an internet corpus; and

a correction generator operative to provide a correction output based at least partially on an evaluation performed by said suspector.

63. A computer-assisted language correction system according to claim 62 and wherein said correction generator includes at least one of missing punctuation

correction functionality, superfluous punctuation correction functionality and punctuation replacement correction functionality.

64. A computer-assisted language correction system comprising:

a grammatical element error suspector evaluating at least some of the words in a language input on the basis of their fit within a context of the language input based on frequency of occurrence of feature-grams of said language input in an internet corpus; and

a correction generator operative to provide a correction output based at least partially on an evaluation performed by said suspector.

65. A computer-assisted language correction system according to claim 64 and wherein said correction generator includes at least one of missing grammatical element correction functionality, superfluous grammatical element correction functionality and grammatical element replacement correction functionality.

66. A computer-assisted language correction system according to claim 64 or claim 65 and wherein said grammatical element is one of an article, a preposition and a conjunction.

FIG. 1

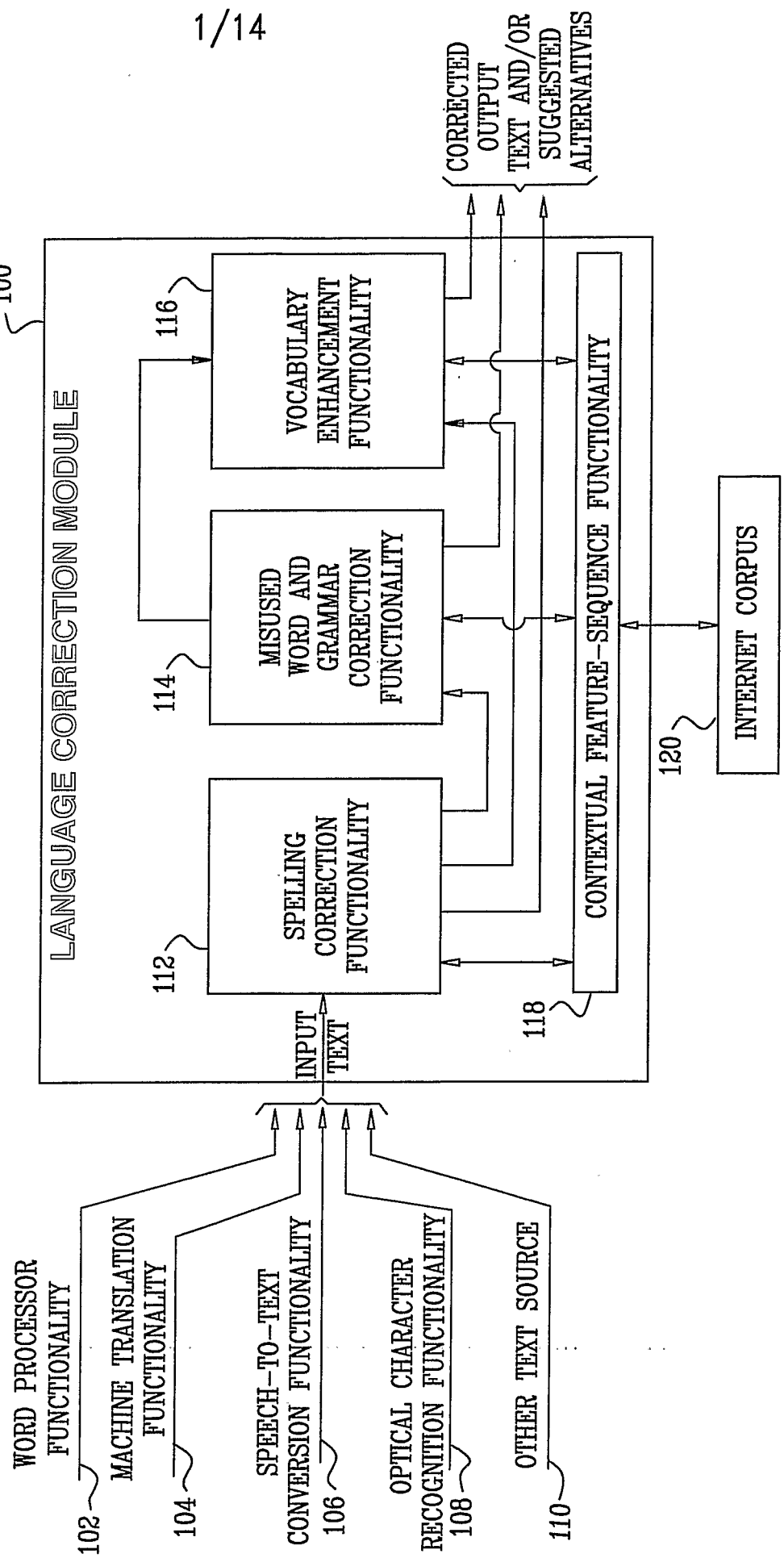

100
LANGUAGE CORRECTION MODULE
INPUT TEXT
112
SPELLING CORRECTION FUNCTIONALITY
114
MISUSED WORD AND GRAMMAR CORRECTION FUNCTIONALITY
116
VOCABULARY ENHANCEMENT FUNCTIONALITY
118
CONTEXTUAL FEATURE-SEQUENCE FUNCTIONALITY
120
INTERNET CORPUS
CORRECTED OUTPUT TEXT AND/OR SUGGESTED ALTERNATIVES
WORD PROCESSOR FUNCTIONALITY
102
MACHINE TRANSLATION FUNCTIONALITY
104
SPEECH-TO-TEXT CONVERSION FUNCTIONALITY
106
OPTICAL CHARACTER RECOGNITION FUNCTIONALITY
108
OTHER TEXT SOURCE
110

FIG. 2

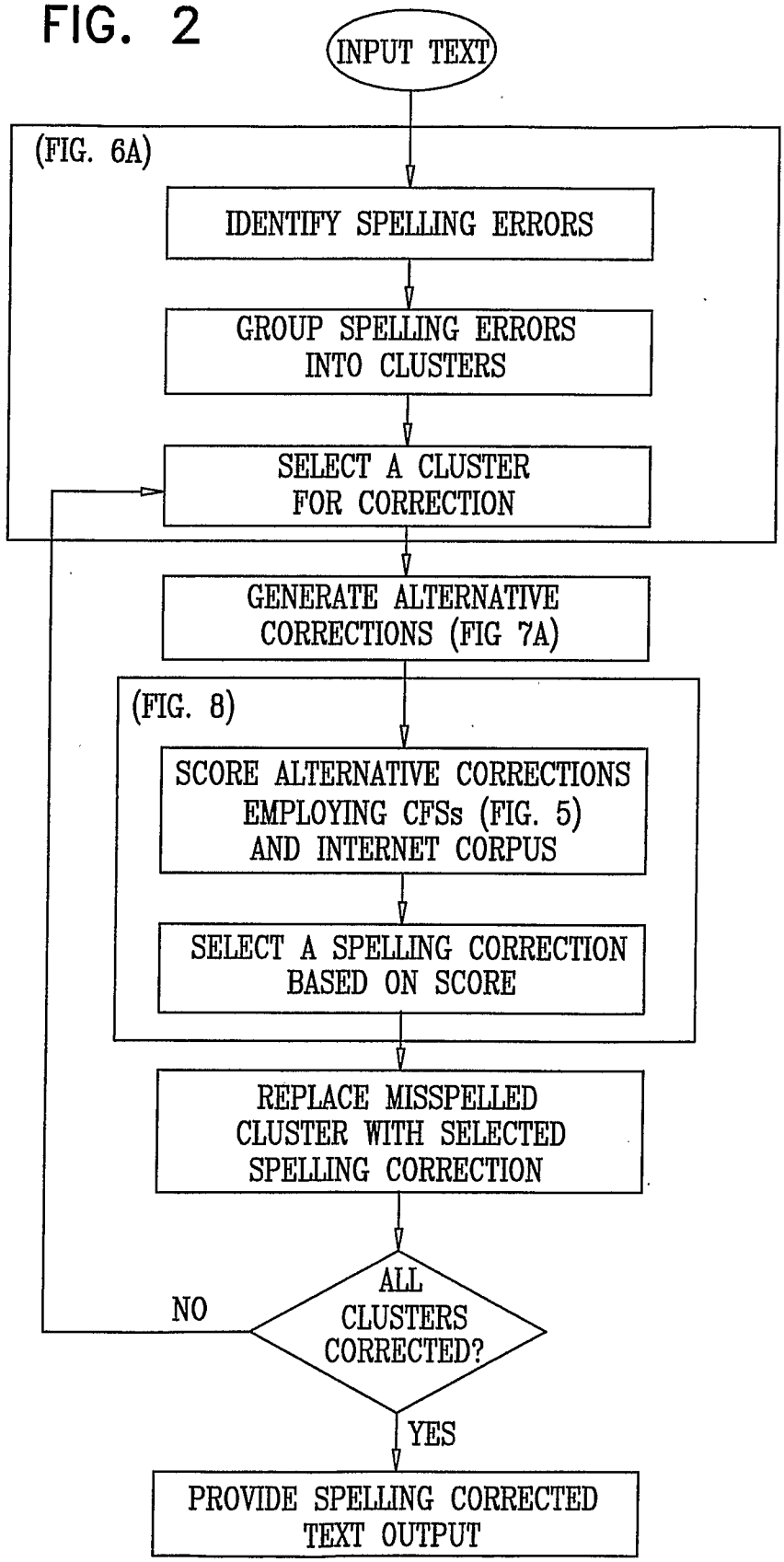
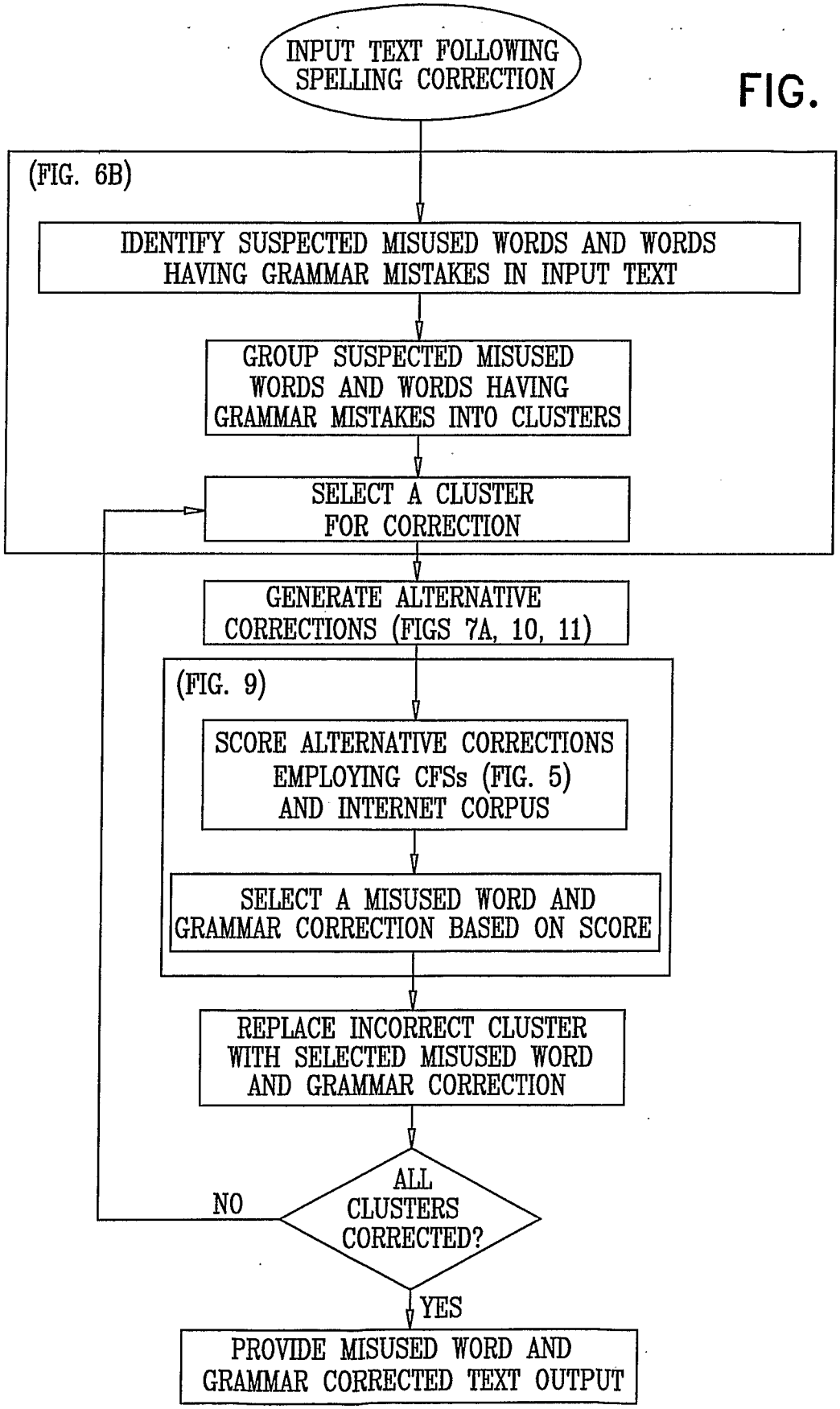

INPUT TEXT
(FIG. 6A)
IDENTIFY SPELLING ERRORS
GROUP SPELLING ERRORS INTO CLUSTERS
SELECT A CLUSTER FOR CORRECTION
GENERATE ALTERNATIVE CORRECTIONS (FIG 7A)
(FIG. 8)
SCORE ALTERNATIVE CORRECTIONS EMPLOYING CFSs (FIG. 5) AND INTERNET CORPUS
SELECT A SPELLING CORRECTION BASED ON SCORE
REPLACE MISSPELLED CLUSTER WITH SELECTED SPELLING CORRECTION
ALL CLUSTERS CORRECTED?
NO
YES
PROVIDE SPELLING CORRECTED TEXT OUTPUT

FIG. 3


INPUT TEXT FOLLOWING SPELLING CORRECTION
(FIG. 6B)
IDENTIFY SUSPECTED MISUSED WORDS AND WORDS HAVING GRAMMAR MISTAKES IN INPUT TEXT
GROUP SUSPECTED MISUSED WORDS AND WORDS HAVING GRAMMAR MISTAKES INTO CLUSTERS
SELECT A CLUSTER FOR CORRECTION
GENERATE ALTERNATIVE CORRECTIONS (FIGS 7A, 10, 11)
(FIG. 9)
SCORE ALTERNATIVE CORRECTIONS EMPLOYING CFSs (FIG. 5) AND INTERNET CORPUS
SELECT A MISUSED WORD AND GRAMMAR CORRECTION BASED ON SCORE
REPLACE INCORRECT CLUSTER WITH SELECTED MISUSED WORD AND GRAMMAR CORRECTION
ALL CLUSTERS CORRECTED?
NO
YES
PROVIDE MISUSED WORD AND GRAMMAR CORRECTED TEXT OUTPUT

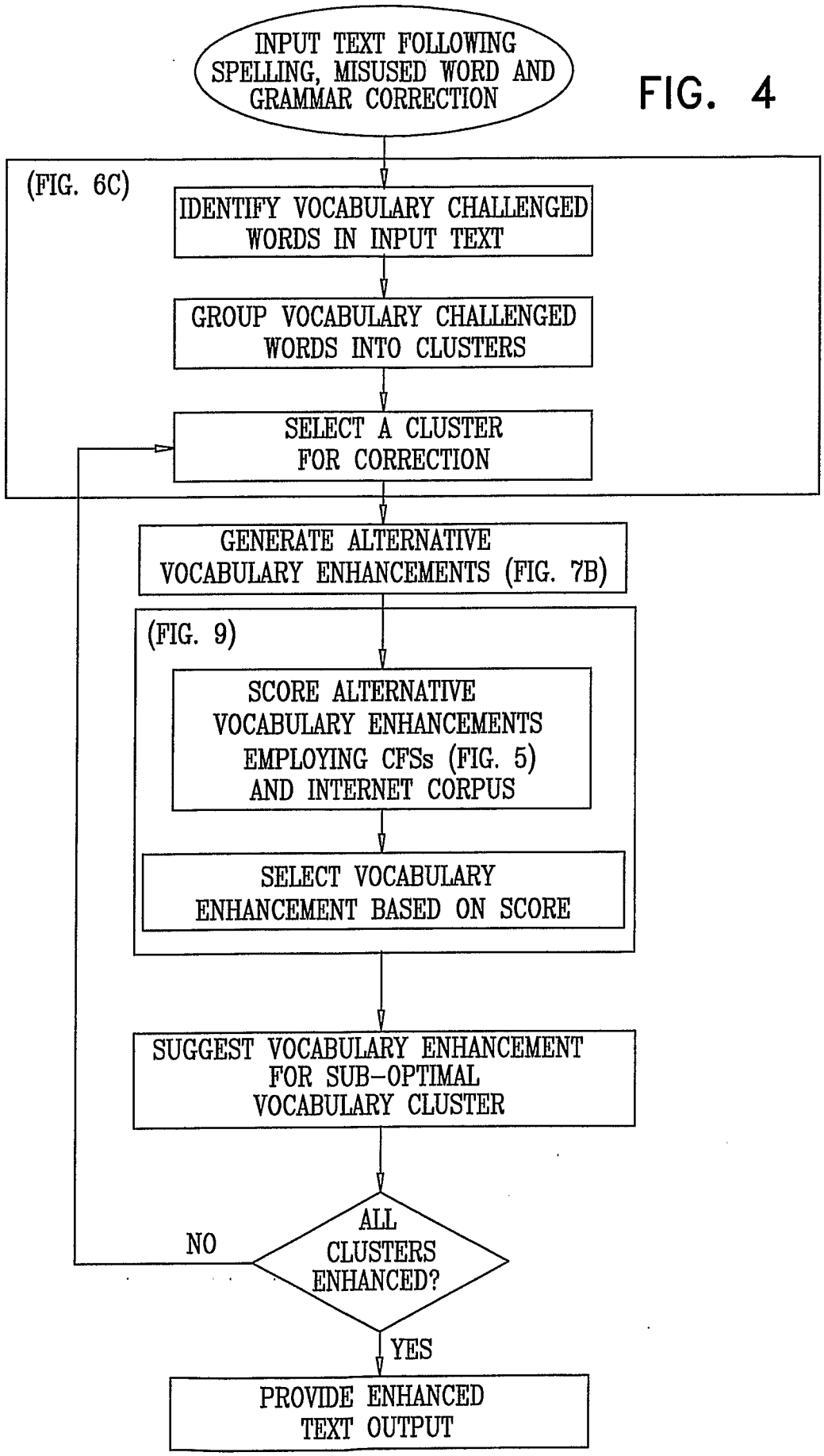

INPUT TEXT FOLLOWING SPELLING, MISUSED WORD AND GRAMMAR CORRECTION
(FIG. 6C)
IDENTIFY VOCABULARY CHALLENGED WORDS IN INPUT TEXT
GROUP VOCABULARY CHALLENGED WORDS INTO CLUSTERS
SELECT A CLUSTER FOR CORRECTION
GENERATE ALTERNATIVE VOCABULARY ENHANCEMENTS (FIG. 7B)
(FIG. 9)
SCORE ALTERNATIVE VOCABULARY ENHANCEMENTS EMPLOYING CFSs (FIG. 5) AND INTERNET CORPUS
SELECT VOCABULARY ENHANCEMENT BASED ON SCORE
SUGGEST VOCABULARY ENHANCEMENT FOR SUB-OPTIMAL VOCABULARY CLUSTER
ALL CLUSTERS ENHANCED?
NO
YES
PROVIDE ENHANCED TEXT OUTPUT

FIG. 4

FIG. 5

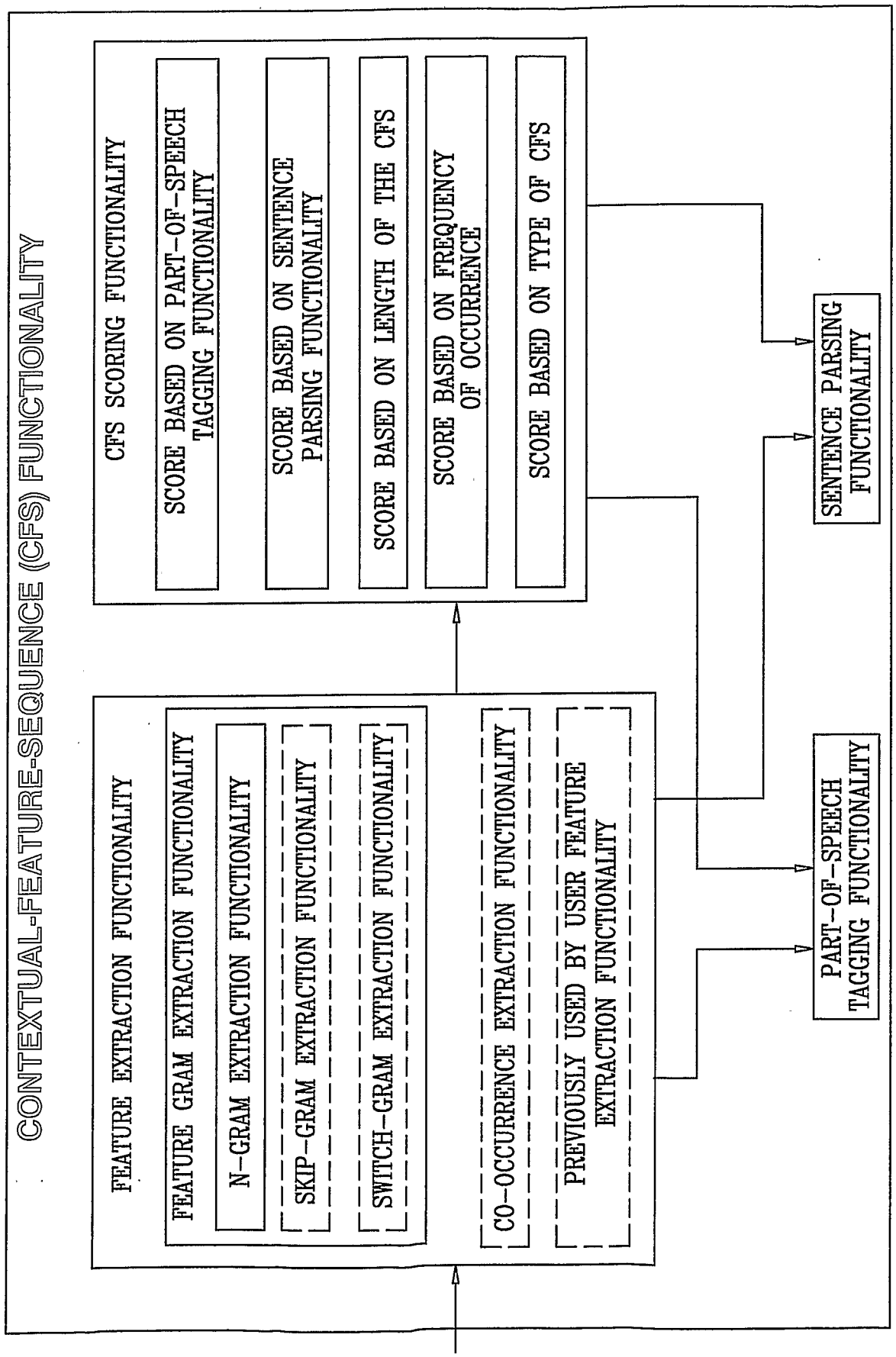
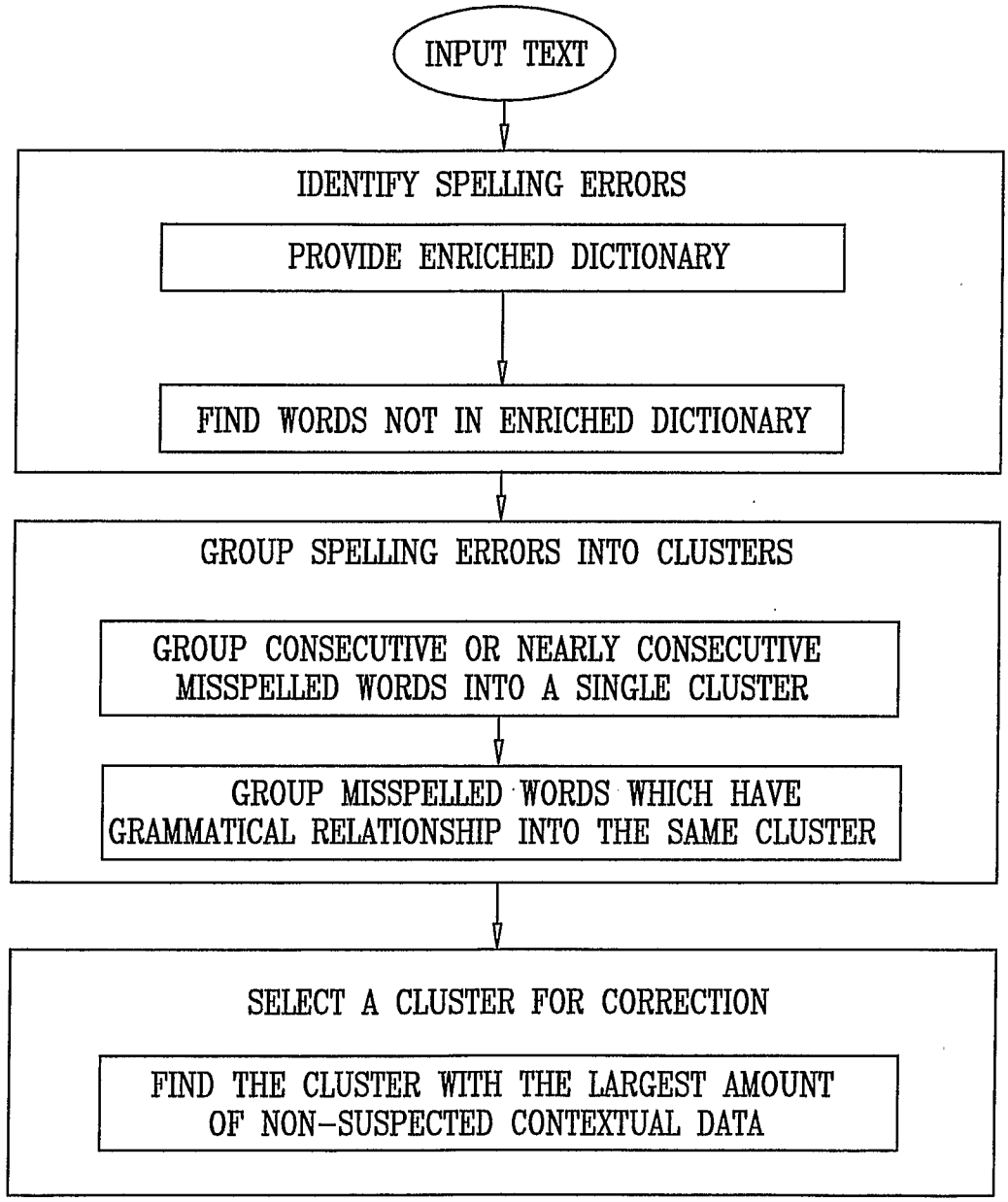

CONTEXTUAL-FEATURE-SEQUENCE (CFS) FUNCTIONALITY
CFS SCORING FUNCTIONALITY
SCORE BASED ON PART-OF-SPEECH TAGGING FUNCTIONALITY
SCORE BASED ON SENTENCE PARSING FUNCTIONALITY
SCORE BASED ON LENGTH OF THE CFS
SCORE BASED ON FREQUENCY OF OCCURRENCE
SCORE BASED ON TYPE OF CFS
FEATURE EXTRACTION FUNCTIONALITY
FEATURE GRAM EXTRACTION FUNCTIONALITY
N-GRAM EXTRACTION FUNCTIONALITY
SKIP-GRAM EXTRACTION FUNCTIONALITY
SWITCH-GRAM EXTRACTION FUNCTIONALITY
CO-OCCURRENCE EXTRACTION FUNCTIONALITY
PREVIOUSLY USED BY USER FEATURE EXTRACTION FUNCTIONALITY
SENTENCE PARSING FUNCTIONALITY
PART-OF-SPEECH TAGGING FUNCTIONALITY

FIG. 6A


INPUT TEXT
IDENTIFY SPELLING ERRORS
PROVIDE ENRICHED DICTIONARY
FIND WORDS NOT IN ENRICHED DICTIONARY
GROUP SPELLING ERRORS INTO CLUSTERS
GROUP CONSECUTIVE OR NEARLY CONSECUTIVE MISSPELLED WORDS INTO A SINGLE CLUSTER
GROUP MISSPELLED WORDS WHICH HAVE GRAMMATICAL RELATIONSHIP INTO THE SAME CLUSTER
SELECT A CLUSTER FOR CORRECTION
FIND THE CLUSTER WITH THE LARGEST AMOUNT OF NON-SUSPECTED CONTEXTUAL DATA

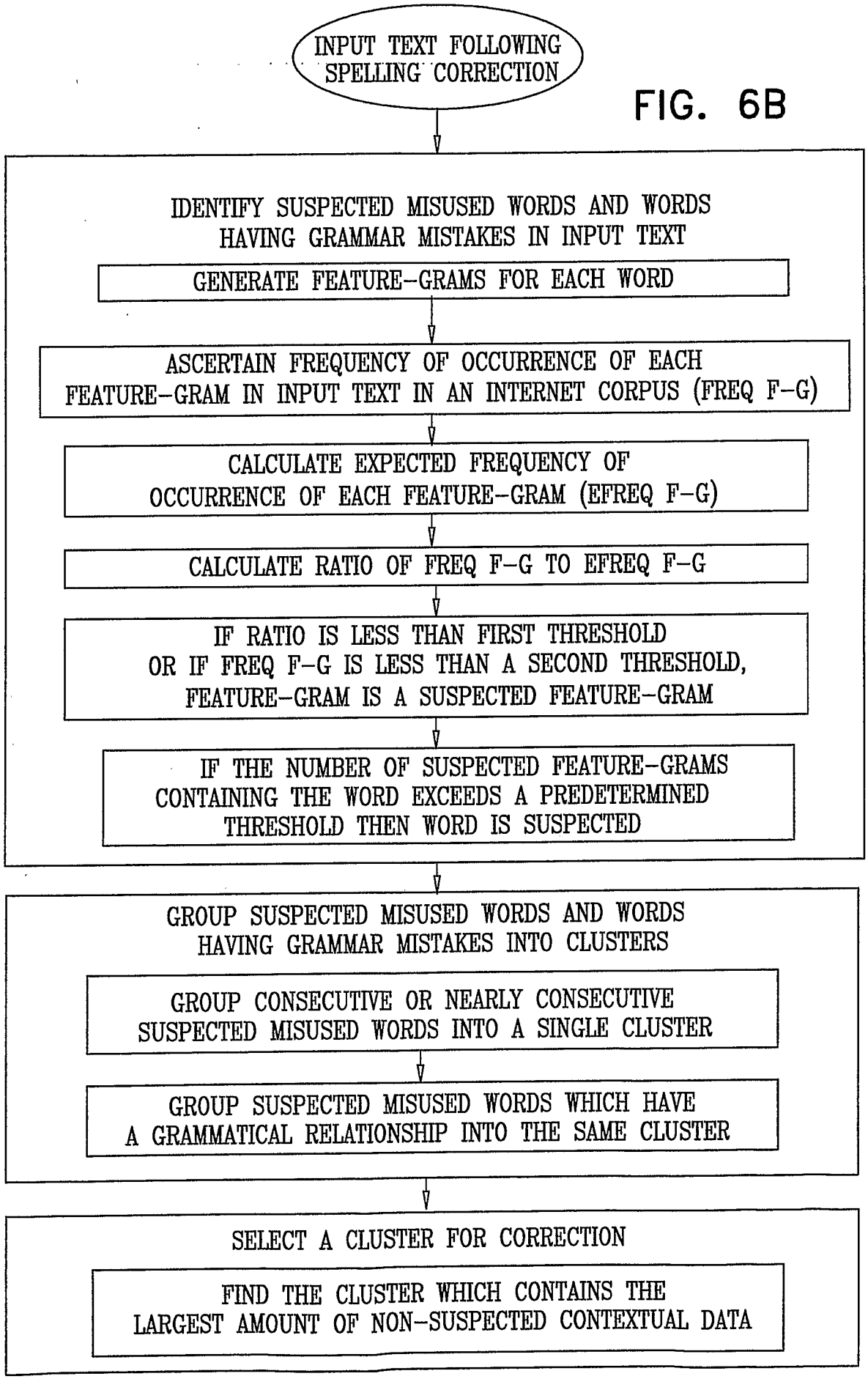

INPUT TEXT FOLLOWING SPELLING CORRECTION
IDENTIFY SUSPECTED MISUSED WORDS AND WORDS HAVING GRAMMAR MISTAKES IN INPUT TEXT
GENERATE FEATURE-GRAMS FOR EACH WORD
ASCERTAIN FREQUENCY OF OCCURRENCE OF EACH FEATURE-GRAM IN INPUT TEXT IN AN INTERNET CORPUS (FREQ F-G)
CALCULATE EXPECTED FREQUENCY OF OCCURRENCE OF EACH FEATURE-GRAM (EFREQ F-G)
CALCULATE RATIO OF FREQ F-G TO EFREQ F-G
IF RATIO IS LESS THAN FIRST THRESHOLD OR IF FREQ F-G IS LESS THAN A SECOND THRESHOLD, FEATURE-GRAM IS A SUSPECTED FEATURE-GRAM
IF THE NUMBER OF SUSPECTED FEATURE-GRAMS CONTAINING THE WORD EXCEEDS A PREDETERMINED THRESHOLD THEN WORD IS SUSPECTED
GROUP SUSPECTED MISUSED WORDS AND WORDS HAVING GRAMMAR MISTAKES INTO CLUSTERS
GROUP CONSECUTIVE OR NEARLY CONSECUTIVE SUSPECTED MISUSED WORDS INTO A SINGLE CLUSTER
GROUP SUSPECTED MISUSED WORDS WHICH HAVE A GRAMMATICAL RELATIONSHIP INTO THE SAME CLUSTER
SELECT A CLUSTER FOR CORRECTION
FIND THE CLUSTER WHICH CONTAINS THE LARGEST AMOUNT OF NON-SUSPECTED CONTEXTUAL DATA

FIG. 6B

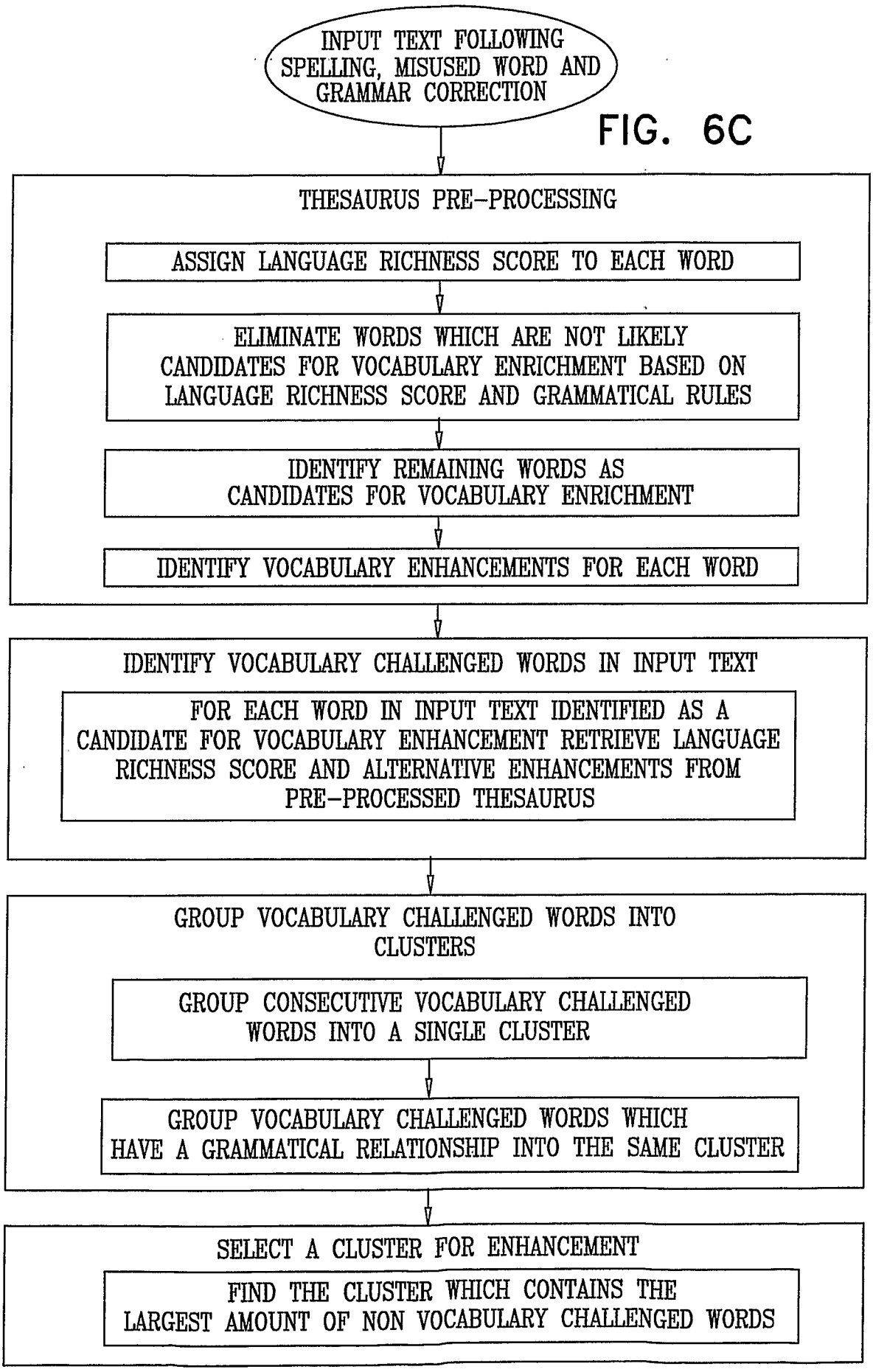

INPUT TEXT FOLLOWING SPELLING, MISUSED WORD AND GRAMMAR CORRECTION
THESAURUS PRE-PROCESSING
ASSIGN LANGUAGE RICHNESS SCORE TO EACH WORD
ELIMINATE WORDS WHICH ARE NOT LIKELY CANDIDATES FOR VOCABULARY ENRICHMENT BASED ON LANGUAGE RICHNESS SCORE AND GRAMMATICAL RULES
IDENTIFY REMAINING WORDS AS CANDIDATES FOR VOCABULARY ENRICHMENT
IDENTIFY VOCABULARY ENHANCEMENTS FOR EACH WORD
IDENTIFY VOCABULARY CHALLENGED WORDS IN INPUT TEXT
FOR EACH WORD IN INPUT TEXT IDENTIFIED AS A CANDIDATE FOR VOCABULARY ENHANCEMENT RETRIEVE LANGUAGE RICHNESS SCORE AND ALTERNATIVE ENHANCEMENTS FROM PRE-PROCESSED THESAURUS
GROUP VOCABULARY CHALLENGED WORDS INTO CLUSTERS
GROUP CONSECUTIVE VOCABULARY CHALLENGED WORDS INTO A SINGLE CLUSTER
GROUP VOCABULARY CHALLENGED WORDS WHICH HAVE A GRAMMATICAL RELATIONSHIP INTO THE SAME CLUSTER
SELECT A CLUSTER FOR ENHANCEMENT
FIND THE CLUSTER WHICH CONTAINS THE LARGEST AMOUNT OF NON VOCABULARY CHALLENGED WORDS

FIG. 6C

FIG. 7A

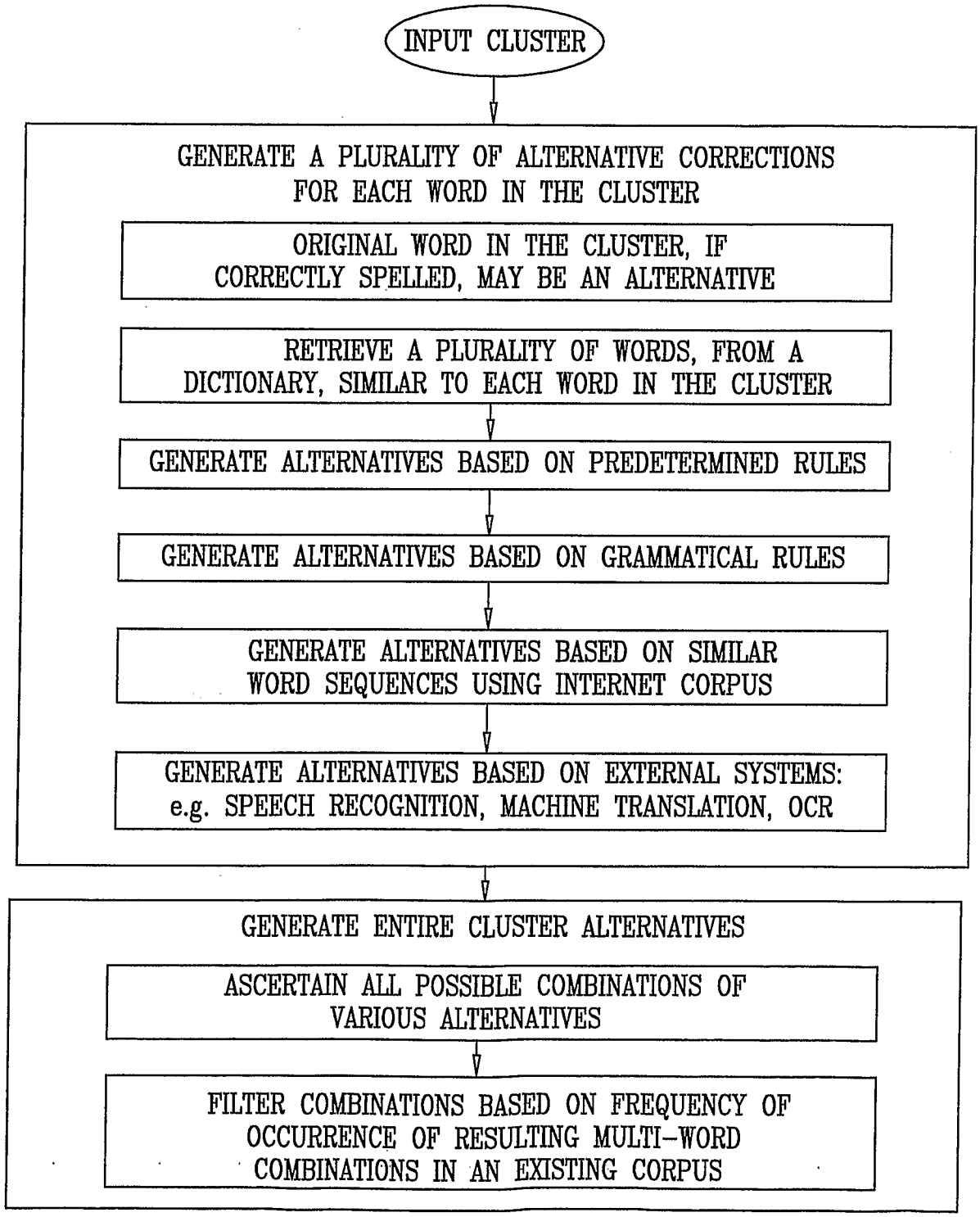

INPUT CLUSTER
GENERATE A PLURALITY OF ALTERNATIVE CORRECTIONS FOR EACH WORD IN THE CLUSTER
ORIGINAL WORD IN THE CLUSTER, IF CORRECTLY SPELLED, MAY BE AN ALTERNATIVE
RETRIEVE A PLURALITY OF WORDS, FROM A DICTIONARY, SIMILAR TO EACH WORD IN THE CLUSTER
GENERATE ALTERNATIVES BASED ON PREDETERMINED RULES
GENERATE ALTERNATIVES BASED ON GRAMMATICAL RULES
GENERATE ALTERNATIVES BASED ON SIMILAR WORD SEQUENCES USING INTERNET CORPUS
GENERATE ALTERNATIVES BASED ON EXTERNAL SYSTEMS: e.g. SPEECH RECOGNITION, MACHINE TRANSLATION, OCR
GENERATE ENTIRE CLUSTER ALTERNATIVES
ASCERTAIN ALL POSSIBLE COMBINATIONS OF VARIOUS ALTERNATIVES
FILTER COMBINATIONS BASED ON FREQUENCY OF OCCURRENCE OF RESULTING MULTI-WORD COMBINATIONS IN AN EXISTING CORPUS

FIG. 7B

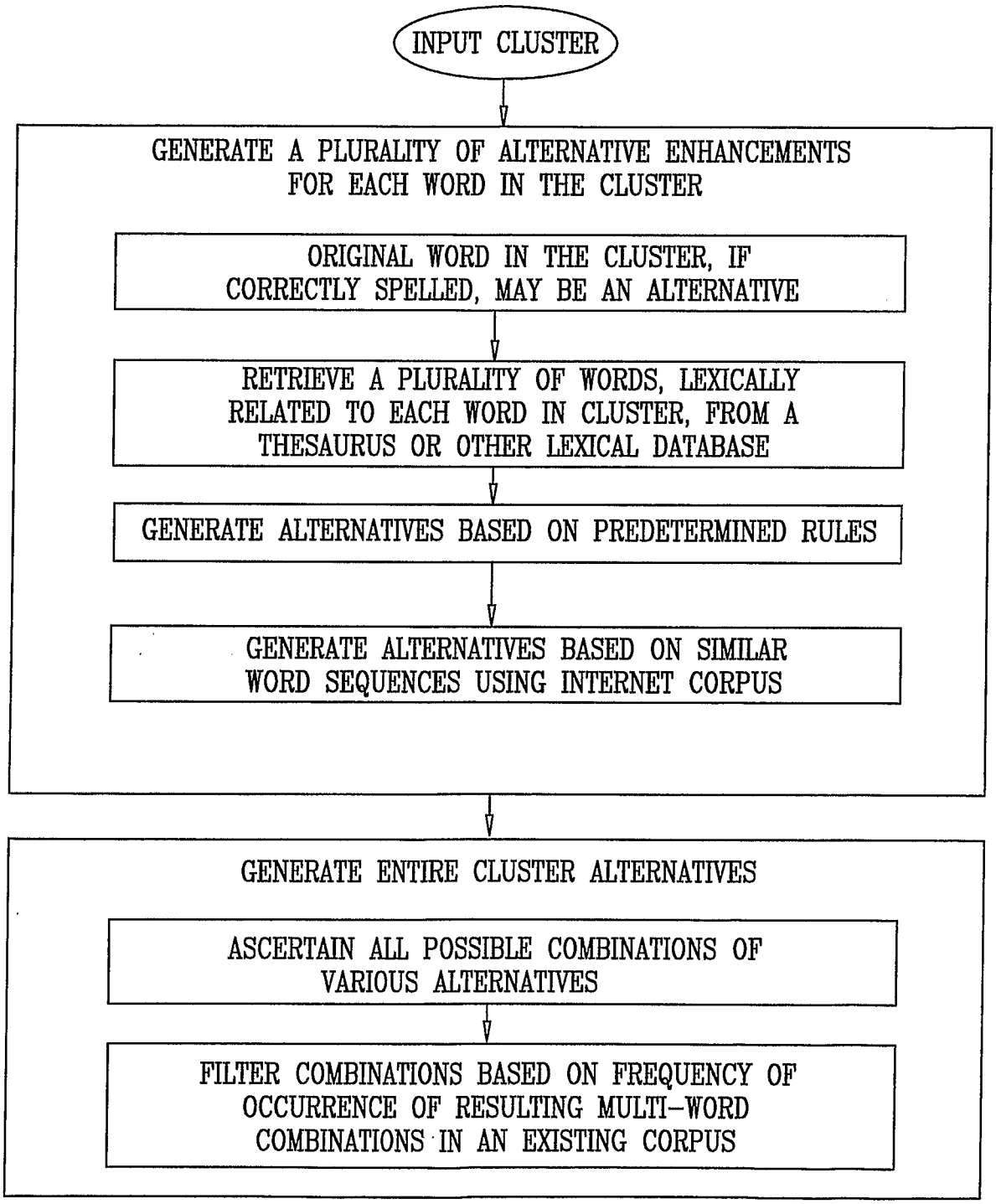

INPUT CLUSTER
GENERATE A PLURALITY OF ALTERNATIVE ENHANCEMENTS FOR EACH WORD IN THE CLUSTER
ORIGINAL WORD IN THE CLUSTER, IF CORRECTLY SPELLED, MAY BE AN ALTERNATIVE
RETRIEVE A PLURALITY OF WORDS, LEXICALLY RELATED TO EACH WORD IN CLUSTER, FROM A THESAURUS OR OTHER LEXICAL DATABASE
GENERATE ALTERNATIVES BASED ON PREDETERMINED RULES
GENERATE ALTERNATIVES BASED ON SIMILAR WORD SEQUENCES USING INTERNET CORPUS
GENERATE ENTIRE CLUSTER ALTERNATIVES
ASCERTAIN ALL POSSIBLE COMBINATIONS OF VARIOUS ALTERNATIVES
FILTER COMBINATIONS BASED ON FREQUENCY OF OCCURRENCE OF RESULTING MULTI-WORD COMBINATIONS IN AN EXISTING CORPUS

FIG. 8

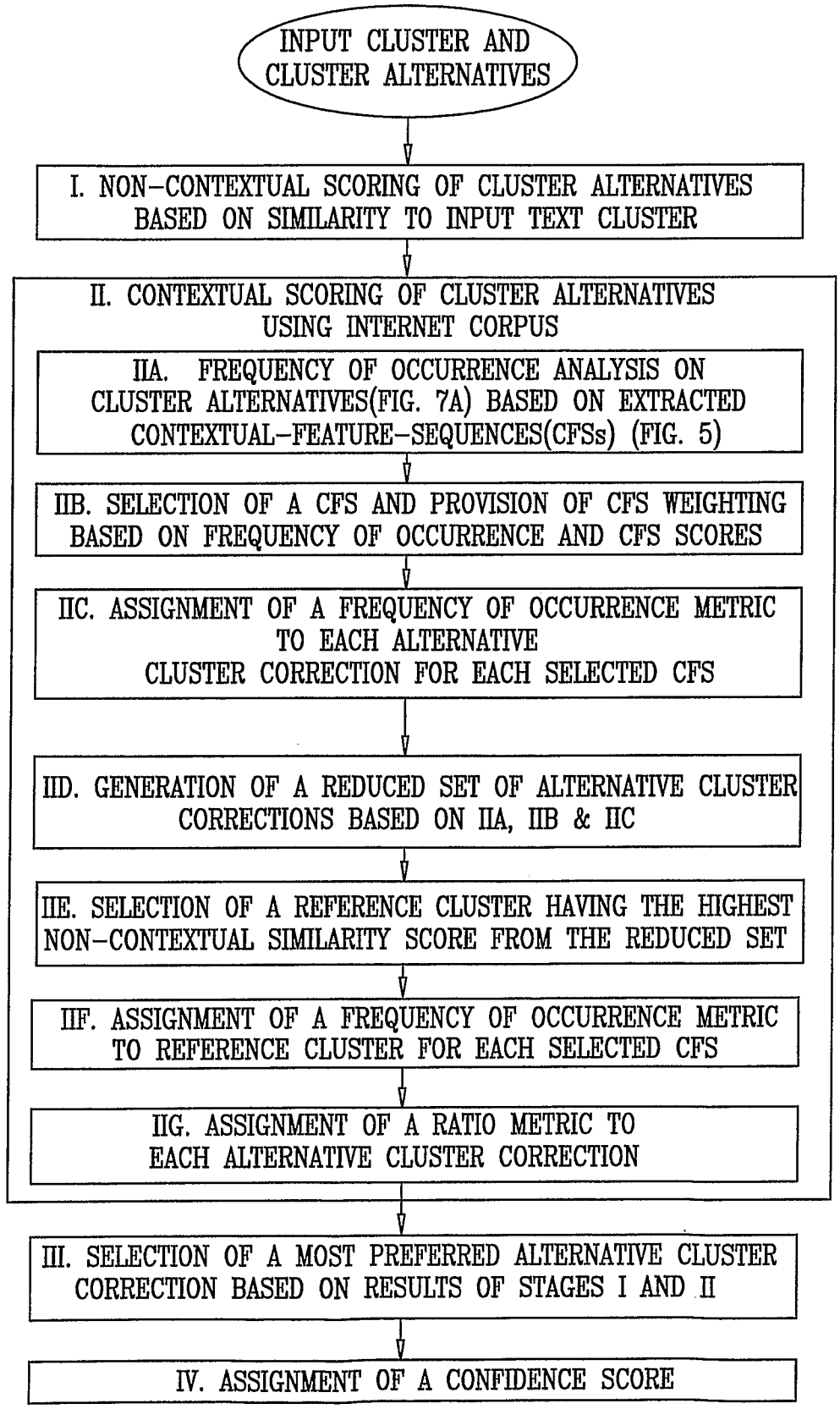

INPUT CLUSTER AND CLUSTER ALTERNATIVES
I. NON-CONTEXTUAL SCORING OF CLUSTER ALTERNATIVES BASED ON SIMILARITY TO INPUT TEXT CLUSTER
II. CONTEXTUAL SCORING OF CLUSTER ALTERNATIVES USING INTERNET CORPUS
IIA. FREQUENCY OF OCCURRENCE ANALYSIS ON CLUSTER ALTERNATIVES(FIG. 7A) BASED ON EXTRACTED CONTEXTUAL-FEATURE-SEQUENCES(CFSs) (FIG. 5)
IIB. SELECTION OF A CFS AND PROVISION OF CFS WEIGHTING BASED ON FREQUENCY OF OCCURRENCE AND CFS SCORES
IIC. ASSIGNMENT OF A FREQUENCY OF OCCURRENCE METRIC TO EACH ALTERNATIVE CLUSTER CORRECTION FOR EACH SELECTED CFS
IID. GENERATION OF A REDUCED SET OF ALTERNATIVE CLUSTER CORRECTIONS BASED ON IIA, IIB & IIC
IIE. SELECTION OF A REFERENCE CLUSTER HAVING THE HIGHEST NON-CONTEXTUAL SIMILARITY SCORE FROM THE REDUCED SET
IIF. ASSIGNMENT OF A FREQUENCY OF OCCURRENCE METRIC TO REFERENCE CLUSTER FOR EACH SELECTED CFS
IIG. ASSIGNMENT OF A RATIO METRIC TO EACH ALTERNATIVE CLUSTER CORRECTION
III. SELECTION OF A MOST PREFERRED ALTERNATIVE CLUSTER CORRECTION BASED ON RESULTS OF STAGES I AND II
IV. ASSIGNMENT OF A CONFIDENCE SCORE

FIG. 9

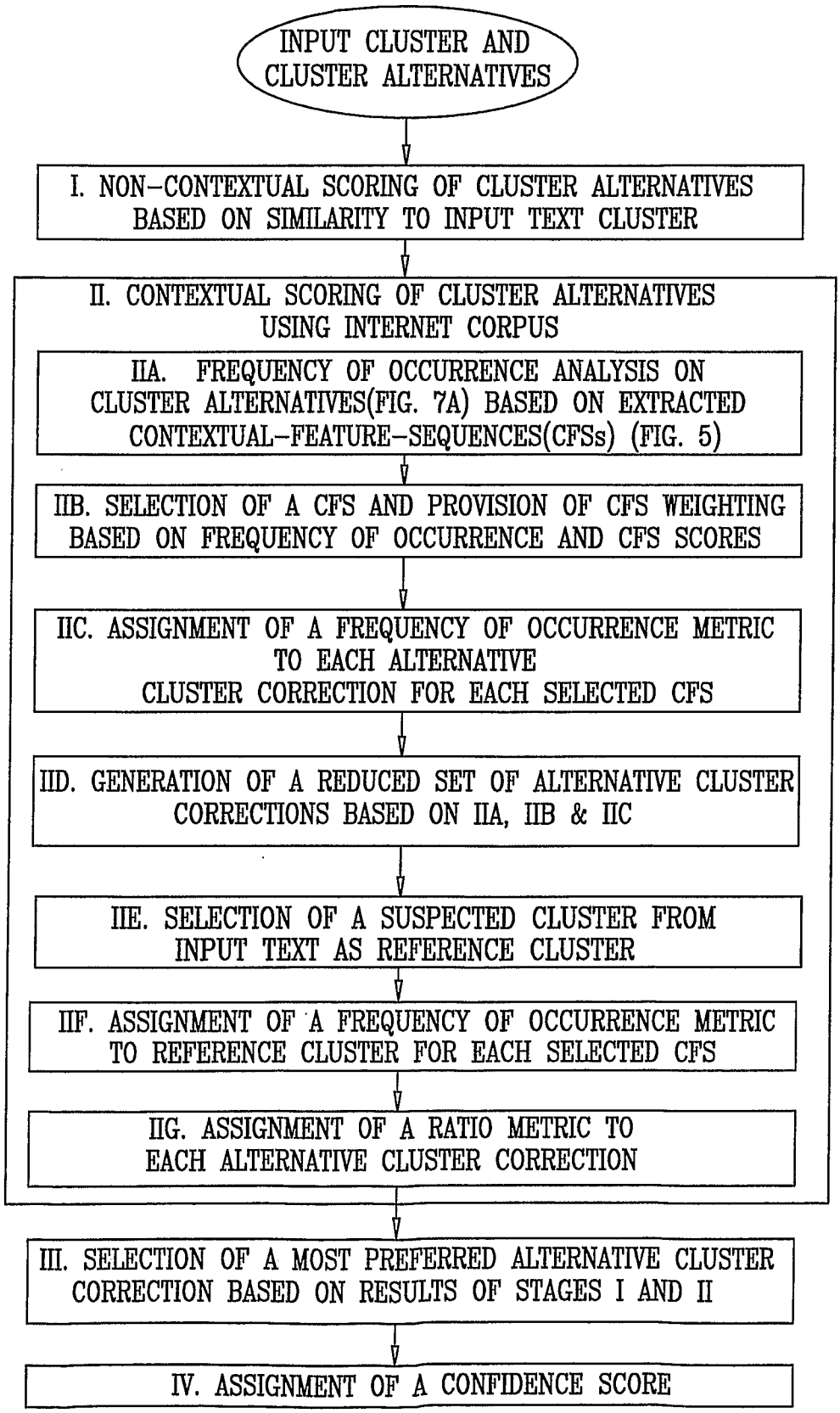

INPUT CLUSTER AND CLUSTER ALTERNATIVES
I. NON-CONTEXTUAL SCORING OF CLUSTER ALTERNATIVES BASED ON SIMILARITY TO INPUT TEXT CLUSTER
II. CONTEXTUAL SCORING OF CLUSTER ALTERNATIVES USING INTERNET CORPUS
IIA. FREQUENCY OF OCCURRENCE ANALYSIS ON CLUSTER ALTERNATIVES(FIG. 7A) BASED ON EXTRACTED CONTEXTUAL-FEATURE-SEQUENCES(CFSs) (FIG. 5)
IIB. SELECTION OF A CFS AND PROVISION OF CFS WEIGHTING BASED ON FREQUENCY OF OCCURRENCE AND CFS SCORES
IIC. ASSIGNMENT OF A FREQUENCY OF OCCURRENCE METRIC TO EACH ALTERNATIVE CLUSTER CORRECTION FOR EACH SELECTED CFS
IID. GENERATION OF A REDUCED SET OF ALTERNATIVE CLUSTER CORRECTIONS BASED ON IIA, IIB & IIC
IIE. SELECTION OF A SUSPECTED CLUSTER FROM INPUT TEXT AS REFERENCE CLUSTER
IIF. ASSIGNMENT OF A FREQUENCY OF OCCURRENCE METRIC TO REFERENCE CLUSTER FOR EACH SELECTED CFS
IIG. ASSIGNMENT OF A RATIO METRIC TO EACH ALTERNATIVE CLUSTER CORRECTION
III. SELECTION OF A MOST PREFERRED ALTERNATIVE CLUSTER CORRECTION BASED ON RESULTS OF STAGES I AND II
IV. ASSIGNMENT OF A CONFIDENCE SCORE

FIG. 10

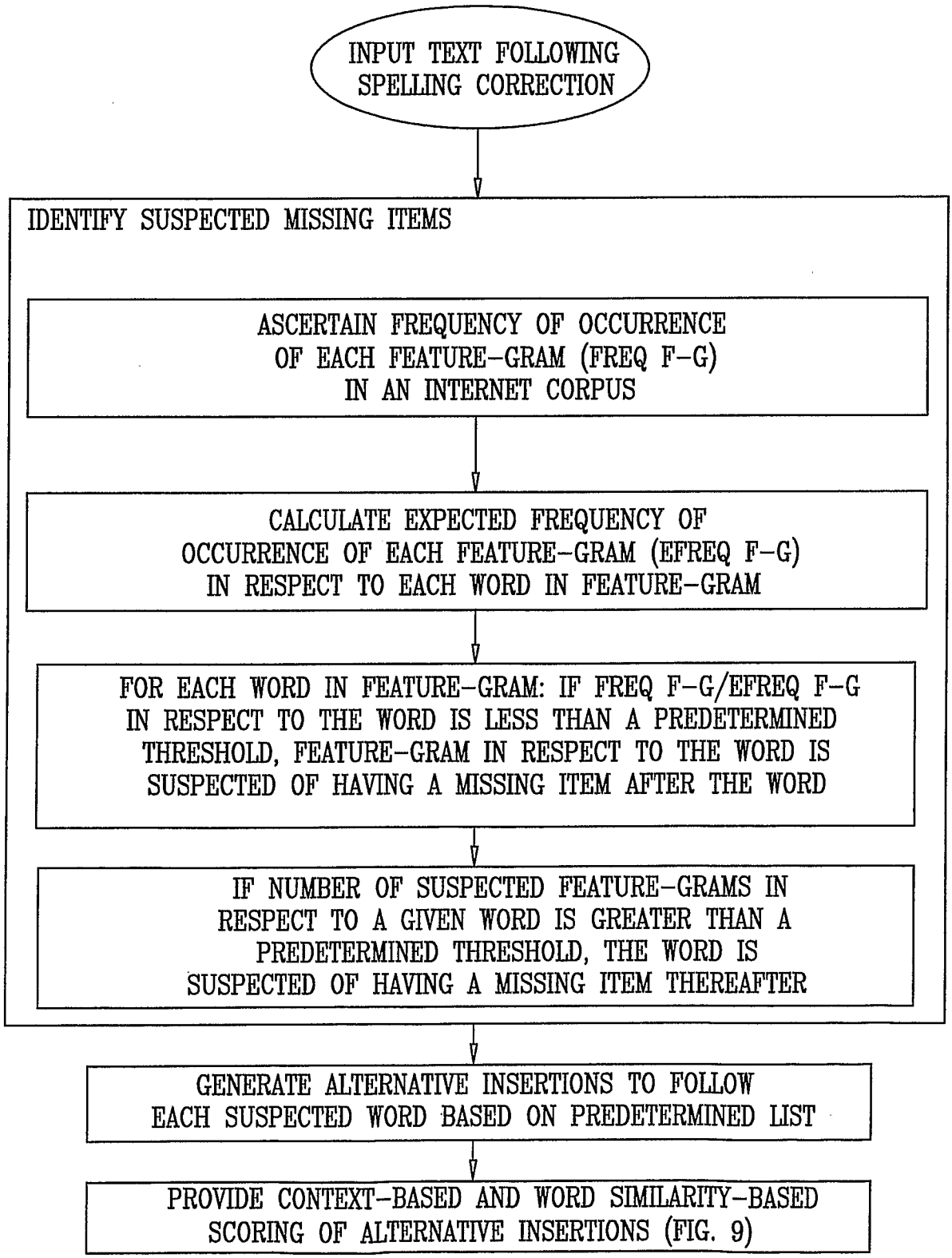
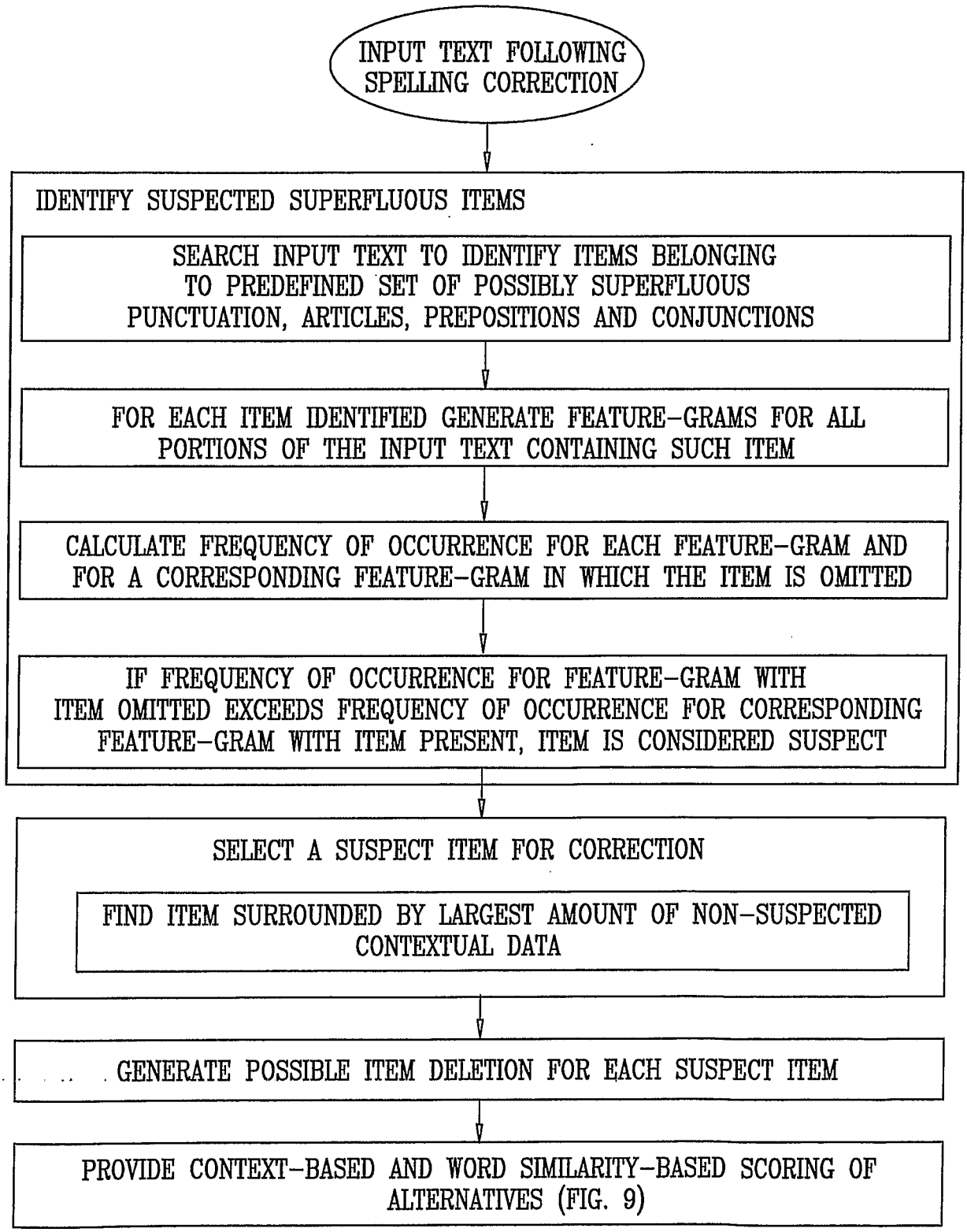

INPUT TEXT FOLLOWING SPELLING CORRECTION
IDENTIFY SUSPECTED MISSING ITEMS
ASCERTAIN FREQUENCY OF OCCURRENCE OF EACH FEATURE-GRAM (FREQ F-G) IN AN INTERNET CORPUS
CALCULATE EXPECTED FREQUENCY OF OCCURRENCE OF EACH FEATURE-GRAM (EFREQ F-G) IN RESPECT TO EACH WORD IN FEATURE-GRAM
FOR EACH WORD IN FEATURE-GRAM: IF FREQ F-G/EFREQ F-G IN RESPECT TO THE WORD IS LESS THAN A PREDETERMINED THRESHOLD, FEATURE-GRAM IN RESPECT TO THE WORD IS SUSPECTED OF HAVING A MISSING ITEM AFTER THE WORD
IF NUMBER OF SUSPECTED FEATURE-GRAMS IN RESPECT TO A GIVEN WORD IS GREATER THAN A PREDETERMINED THRESHOLD, THE WORD IS SUSPECTED OF HAVING A MISSING ITEM THEREAFTER
GENERATE ALTERNATIVE INSERTIONS TO FOLLOW EACH SUSPECTED WORD BASED ON PREDETERMINED LIST
PROVIDE CONTEXT-BASED AND WORD SIMILARITY-BASED SCORING OF ALTERNATIVE INSERTIONS (FIG. 9)

FIG. 11


INPUT TEXT FOLLOWING SPELLING CORRECTION
IDENTIFY SUSPECTED SUPERFLUOUS ITEMS
SEARCH INPUT TEXT TO IDENTIFY ITEMS BELONGING TO PREDEFINED SET OF POSSIBLY SUPERFLUOUS PUNCTUATION, ARTICLES, PREPOSITIONS AND CONJUNCTIONS
FOR EACH ITEM IDENTIFIED GENERATE FEATURE-GRAMS FOR ALL PORTIONS OF THE INPUT TEXT CONTAINING SUCH ITEM
CALCULATE FREQUENCY OF OCCURRENCE FOR EACH FEATURE-GRAM AND FOR A CORRESPONDING FEATURE-GRAM IN WHICH THE ITEM IS OMITTED
IF FREQUENCY OF OCCURRENCE FOR FEATURE-GRAM WITH ITEM OMITTED EXCEEDS FREQUENCY OF OCCURRENCE FOR CORRESPONDING FEATURE-GRAM WITH ITEM PRESENT, ITEM IS CONSIDERED SUSPECT
SELECT A SUSPECT ITEM FOR CORRECTION
FIND ITEM SURROUNDED BY LARGEST AMOUNT OF NON-SUSPECTED CONTEXTUAL DATA
GENERATE POSSIBLE ITEM DELETION FOR EACH SUSPECT ITEM
PROVIDE CONTEXT-BASED AND WORD SIMILARITY-BASED SCORING OF ALTERNATIVES (FIG. 9)