

12

DEMANDE DE BREVET D'INVENTION

A1

22 Date de dépôt : 08.01.07.

30 Priorité :

43 Date de mise à la disposition du public de la
demande : 11.07.08 Bulletin 08/28.

56 Liste des documents cités dans le rapport de
recherche préliminaire : *Se reporter à la fin du
présent fascicule*

60 Références à d'autres documents nationaux
apparentés :

71 Demandeur(s) : SAGEM COMMUNICATION Société
anonyme — FR.

72 Inventeur(s) : BERGER JEROME.

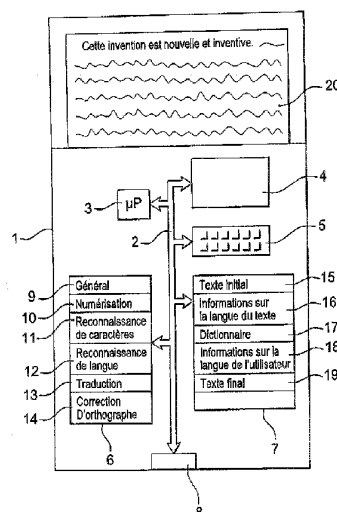
73 Titulaire(s) :

74 Mandataire(s) : CABINET SCHMIT CHRETIEN
SCHIHN.

54 PROCÉDE D'ÉDITION D'UN TEXTE EXPRIME DANS UNE LANGUE.

57 Dans l'invention, on propose de minimiser la charge
globale de calcul et d'assouplir l'analyse d'un texte écrit (15)
vis-à-vis des erreurs dues notamment à des mauvaises re-
connaissance (11) de caractères et/ou à des fautes de frap-
pe.

A cet effet, dans l'invention, on reconnaît (12) la langue
d'écriture du texte écrit en établissant une distribution de
fréquences d'apparition de groupes de lettres, ou n-gram-
mes, dans ce texte, n étant un nombre entier positif non nul,
et en la comparant avec des distributions de référence des
fréquences d'apparition des n-grammes dans plusieurs lan-
gues (α ; β ; γ) pour déterminer celle qui est la plus proche.
Plus précisément, cette comparaison est réalisée en calcul-
ant, n-gramme par n-gramme, un écart entre une probabili-
té de présence du n-gramme dans la langue et une
statistique de présence du même n-gramme dans le texte,
pour définir une éventuelle similarité. Ainsi, une faute de
frappe et/ou une mauvaise reconnaissance de caractère
manuscrit, engendrant au moins un n-gramme erroné, sont
négligeables au regard de la quantité de n-grammes correc-
tement distribués.



Procédé d'édition d'un texte exprimé dans une langue

DOMAINE TECHNIQUE DE L'INVENTION

La présente invention a pour objet un procédé d'édition d'un document
5 écrit : un texte.

Un but essentiel du procédé selon l'invention est de reconnaître, de façon automatique et systématique la langue de rédaction du texte, même si ce texte comporte des erreurs liées, par exemple, à des fautes de frappes ou des mauvaises reconnaissances de caractères d'un texte manuscrit.

10 Le domaine de l'invention est, d'une façon générale, celui du traitement de texte. Dans ce domaine, on connaît notamment la correction orthographique et la traduction qui font appel à des dictionnaires. A titre de traitement, on connaît aussi la reconnaissance de caractères manuscrits ou la reconnaissance vocale qui nécessitent l'utilisation d'une base de données
15 relatives aux lettres de différents alphabets. Pour qu'un logiciel de traitement de texte et/ou de reconnaissance de caractères manuscrits ou vocale soit efficace, il faut qu'il comporte un module de dictionnaire qui corrige les erreurs selon une langue de référence. Un des moyens de l'invention est de déterminer automatiquement, par une machine, le dictionnaire qui
20 correspond à un texte donné.

ARRIERE-PLAN TECHNOLOGIQUE DE L'INVENTION

Une première méthode de recherche de dictionnaire comporte la vérification de ce que tous les mots du texte sont présents dans un même dictionnaire. Cette méthode, outre qu'elle impose une grande capacité
25 mémoire (autant de fois qu'il y a de langues à tester), ne résout pas le problème des fautes d'orthographe.

Pour chercher parmi tous les dictionnaires possibles celui qui convient le mieux, dans l'état de la technique, il est aussi connu d'utiliser une approche sémantique. Avec une telle approche, on compte dans le texte des
30 fréquences d'apparition de mots. On compare ces fréquences d'apparition de mots avec des fréquences d'apparition de ces mêmes mots dans un ensemble de textes de la langue. Si la comparaison est favorable, on décide de considérer que le texte est exprimé dans cette langue.

Une contrainte existe également pour cette deuxième méthode : il est
35 nécessaire de disposer d'un dictionnaire conséquent, d'environ 40 000 mots,

2

pour chaque langue. De plus, en termes de calculs, dans les environnements embarqués, cette méthode est trop lourde.

Une autre méthode, décrite dans le document US/1006 002 5988, consiste à utiliser des fréquences d'apparition de groupes ou séquences de n lettres appelés n-grammes, n étant un nombre entier positif non nul. Ce procédé permet d'utiliser un nombre beaucoup moins important de données. Par exemple, pour un alphabet de 26 lettres, on aura :

- $26^3 = 17\,576$ combinaisons pour des trigrammes ;

- $26^2 = 676$ combinaisons pour des bigrammes.

Selon ce document, on obtient un profil statistique représentant la probabilité pour chaque n-gramme d'appartenir à une langue donnée. Ce profil statistique est ensuite traité en multipliant chacune de ses probabilités entre elles, de façon à calculer une probabilité globale. On calcule la probabilité globale pour plusieurs langues candidates. On choisit ensuite la probabilité globale la plus élevée pour déterminer la langue candidate qui sera considérée comme étant celle du texte.

Avec cet état de la technique, un problème demeure. Si le texte étudié comporte un n-gramme n'existant pas dans une langue candidate, par exemple "zz" dans le cas d'un bigramme, on a pour ce n-gramme une probabilité nulle d'appartenance à ladite langue. De ce fait, la probabilité globale est nulle également car elle est le résultat de la multiplication des probabilités propres à chaque n-gramme entre elles. Un tel n-gramme erroné, pouvant être provoqué notamment par une faute de frappe ou par une mauvaise reconnaissance de caractères, empêche la reconnaissance d'une langue.

De plus, tous les n-grammes possibles ne sont pas forcément présents dans le texte étudié et/ou certains n-grammes y sont plusieurs fois. Pour ceux qui y sont plusieurs fois, leur présence est prise en compte plusieurs fois. Cette dernière méthode présente un manque de souplesse vis-à-vis des erreurs dues notamment à des mauvaises reconnaissances de caractères ou à des fautes de frappe.

Les fréquences d'apparition de n-grammes sont déjà utilisées dans le domaine du traitement du langage, en particulier pour la reconnaissance vocale. Une telle méthode de reconnaissance vocale consiste à analyser des fréquences d'apparition de n-grammes pour choisir un mot parmi d'autres.

3

Cette méthode de reconnaissance vocale ne sert pas pour reconnaître la langue utilisée. En particulier, un procédé de reconnaissance vocale nécessite a priori la connaissance de la langue dans laquelle le locuteur s'exprime. Il ne peut être question de la rechercher.

5 DESCRIPTION GENERALE DE L'INVENTION

L'objet de l'invention propose une solution au problème qui vient d'être exposé. Dans l'invention, on propose de minimiser la charge globale de calcul et d'assouplir l'analyse d'un texte écrit vis-à-vis des erreurs dues notamment à des mauvaises reconnaissance de caractères et/ou à des
10 fautes de frappe.

A cet effet, dans l'invention, on propose de reconnaître la langue d'écriture d'un texte écrit en établissant des statistiques relatives à une distribution de fréquences d'apparition de groupes de lettres, ou n-grammes, dans ce texte, n étant un nombre entier positif non nul, et en la comparant
15 avec des statistiques de référence relatives à des distribution de fréquences d'apparition des n-grammes dans plusieurs langues pour déterminer celle qui est la plus proche.

Plus précisément, la comparaison de la distribution du texte avec les distributions respectives des langues prédéfinies est réalisée en calculant, n-
20 gramme par n-gramme, un écart entre une probabilité de présence du n-gramme dans la langue et une statistique de présence du même n-gramme dans le texte.

Dans un exemple, la somme de toutes les valeurs absolues des écarts séparant les n-grammes sert pour définir une éventuelle similarité. D'autres
25 méthodes mathématiques existent pour quantifier ces écarts, à savoir la somme des carrés desdits écarts, ou le maximum des valeurs absolues desdits écarts ou autres. Ainsi, une faute de frappe et/ou une mauvaise reconnaissance de caractère manuscrit, engendrant au moins un n-gramme erroné, sont négligeables au regard de la quantité de n-grammes
30 correctement distribués. Par ailleurs, ce procédé représente une charge de travail moins importante que dans l'état de la technique.

L'invention concerne donc essentiellement un procédé d'édition d'un texte exprimé dans une langue, comportant des étapes dans lesquelles :

- on reçoit, dans un appareil, une suite d'informations binaires
35 correspondant à des caractères d'un alphabet ;

4

- on édite, sur un périphérique de l'appareil, un texte correspondant aux caractères reçus ;

- on fragmente le texte en une pluralité de groupes de n caractères appelés n -grammes, n étant un nombre entier positif non nul ;

5 - on établit une distribution texte de fréquences auxquelles apparaissent les n -grammes dans le texte ;

- on établit une distribution langue de fréquences auxquelles apparaissent les n -grammes dans une langue ;

10 - on réitère cette dernière opération pour un certain nombre de langues connues, chaque langue connue étant caractérisée par une distribution de fréquences d'apparition de n -grammes ;

- on compare lesdites distributions texte et langues entre elles ;

- on détermine comme langue du texte, la langue dont la distribution langue présente une plus grande similarité avec la distribution texte et on
15 traite le texte en fonction de cette détermination ;

caractérisé en ce qu'il comporte l'étapes suivante :

- on calcule, n -gramme par n -gramme, l'écart entre la fréquence d'apparition d'un n -gramme dans le texte et dans la langue.

20 Outre les caractéristiques principales qui viennent d'être mentionnées dans le paragraphe précédent, le procédé selon l'invention peut présenter une ou plusieurs caractéristiques complémentaires parmi les suivantes:

- on élargit les combinaisons possibles réalisées avec l'ensemble des lettres de l'alphabet de la langue avec un ou deux caractères d'espace placés avant ou après un $(n-1)$ gramme ou avant et après un $(n-2)$ gramme ;

25 - les n -grammes sont des groupes de deux caractères appelés bigrammes ;

- les n -grammes sont des groupes de trois caractères appelés trigrammes ;

30 - la comparaison des distributions comporte une étape d'addition, pour tous les n -grammes, de toutes les valeurs absolues des écarts les séparant ;

- la comparaison des distributions comporte une étape d'addition, pour tous les n -grammes, de tous les carrés des écarts les séparant ;

35 - la comparaison des distributions comporte une étape de détermination du maximum de toutes les valeurs absolues des écarts les séparant ;

5

- on ne retient les écarts que s'ils souscrivent à un certain critère ;
 - le critère auquel doivent souscrire les écarts est de présenter une valeur inférieure à un seuil ;
 - le critère auquel doivent souscrire les écarts est de présenter une
- 5 valeur minimale ;
- le traitement réalisé par le procédé est une correction d'orthographe ;
 - le traitement réalisé par le procédé est une traduction automatique du texte reçu, dans une langue paramétrée par l'appareil.

10 L'invention et ses différentes applications seront mieux comprises à la lecture de la description qui suit et à l'examen des figures qui l'accompagnent.

BREVE DESCRIPTION DES FIGURES

Celles-ci ne sont présentées qu'à titre indicatif et nullement limitatif de l'invention. Les figures montrent :

- 15 - à la figure 1, un exemple de mise en œuvre du procédé selon l'invention dans un terminal ;
- à la figure 2, un exemple de comparaison d'une distribution de bigrammes d'un texte avec des distributions de n-grammes de trois langues distinctes.

20 DESCRIPTION DETAILLEE DES FORMES DE REALISATION PREFEREES DE L'INVENTION

Les différents éléments apparaissant sur plusieurs figures auront gardé, sauf précision contraire, la même référence.

25 La figure 1 montre un exemple de mise en œuvre du procédé selon l'invention dans un terminal. Dans un exemple préféré, le terminal est une imprimante multi-fonctions 1. En variante, l'invention peut être mise en œuvre dans un téléphone mobile. L'invention est particulièrement utile si elle est appliquée à un appareil mobile. La première utilisation envisagée est celle des imprimantes multi-fonctions 1 dans le cadre d'une numérisation

30 avec reconnaissance de caractères. Une telle numérisation est réalisée lorsque l'imprimante multi-fonctions 1 possède un lecteur optique pour produire une image numérique d'un texte posé sur une glace de lecture, ou défilant devant une barre de lecture, de ce lecteur optique. L'image numérique est ensuite convertie en texte par un logiciel de reconnaissance

35 de caractères. Il peut convenir de rechercher automatiquement la langue du

6

texte.

L'imprimante multi-fonctions 1 comporte, reliés par un bus 2 d'adresses, de données et de commandes : un microprocesseur 3, un écran 4, un clavier 5, une mémoire de programmes 6, une mémoire de données 7 et une interface 8.

La mémoire de programmes 6 comprend, à cet effet, éventuellement un programme général 9, éventuellement un programme 10 de numérisation, éventuellement un programme 11 de reconnaissance de caractères, selon l'invention, un programme 12 de reconnaissance de langue, éventuellement un programme de traduction 13 et éventuellement un programme de correction orthographique 14. Dans un exemple, le programme 9 est le système d'exploitation de l'imprimante multi-fonctions 1.

La mémoire de données 7 comprend : un texte initial 15, des informations 16 sur la langue du texte initial, un dictionnaire 17, des informations sur une langue 18 de l'utilisateur et un texte final 19. Les informations 16 sont les informations recherchées.

Selon l'invention, un texte manuscrit 20 est reçu par l'imprimante multi-fonctions 1, par exemple sous la forme d'un document papier comportant des caractères d'un alphabet. Dans un exemple, l'alphabet utilisé est l'alphabet latin qui contient 26 lettres. Le texte manuscrit 20 est numérisé par le programme 10 de numérisation de l'imprimante multi-fonctions 1 sous forme d'une image numérique. Les caractères numérisés sont ensuite reconnus par le programme 11 de reconnaissance de caractères et convertis pour obtenir le texte initial 15. En variante, le texte initial 15 est reçu par une interface 8 reliée au bus 2, voire par un lecteur de type lecteur de carte à mémoire, également relié au bus 2 et recevant une mémoire amovible chargée avec le texte à éditer.

Il existe de nombreux types de carte amovible à mémoire, en général associée chacune avec un protocole d'échange propriétaire. Par exemple, la mémoire amovible peut être un support à puce comportant un mode d'emploi de l'imprimante multi-fonctions 1.

Quand le texte initial 15 est reçu par l'imprimante multi-fonctions 1, il peut être reçu sous la forme d'un document papier comportant des caractères manuscrits d'un alphabet. La langue d'écriture du texte 15 doit être reconnue par le programme 12 de reconnaissance de langue, au moyen

7

du procédé selon l'invention. Dans tous les cas, l'appareil 1 reçoit une suite d'informations binaires correspondant à des caractères d'un alphabet.

Une fois les caractères manuscrits et la langue d'écriture reconnus, voire également avant reconnaissance de la langue, le texte reçu 15 est édité sur un périphérique de l'appareil, voire imprimé sur papier. En variante, l'édition peut être vocale. Dans ce cas, le périphérique est un haut-parleur. Toutefois, au sens de l'invention, il est possible de ne pas éditer le texte reçu tel quel ou de ne l'éditer qu'une fois traité (expurgé les fautes d'orthographe par exemple). Au sens littéral, dans l'invention, on considère en outre que 10 l'édition est réalisée du seul fait que le texte est au moins mémorisé dans la mémoire 7, qu'il soit visualisé, imprimé, ou diffusé ultérieurement, ou immédiatement ou non.

Le texte édité de l'invention correspondant aux caractères reçus est ainsi soit le texte initial 15, soit le texte traité 19.

15 Le programme 12 de reconnaissance de la langue d'écriture du texte 15 comporte les étapes dans lesquelles :

- on fragmente le texte initial 15 en une pluralité de groupes de n caractères appelés n -grammes. Dans un exemple préféré, on prend $n = 2$, i.e. on utilise des bigrammes. En variante, on prend $n = 3$, i.e. on utilise des 20 trigrammes. Les bigrammes offrent l'avantage d'augmenter la vitesse de traitement ;

- on établit une distribution texte de fréquences auxquelles apparaissent les bigrammes dans le texte ;

- on établit une distribution langue de fréquences auxquelles 25 apparaissent les bigrammes dans une langue α ;

- on réitère cette dernière opération pour un certain nombre de langues connues β et γ , chaque langue connue β et γ étant caractérisée par une distribution de fréquences d'apparition de bigrammes ;

- on compare lesdites distributions texte et langues entre elles en 30 calculant n -gramme par n -gramme, bigramme par bigramme dans l'exemple, l'écart entre la fréquence d'apparition d'un n -gramme dans le texte et dans la langue α , β et γ et en additionnant toutes les valeurs absolues desdits écarts. D'autres méthodes mathématiques existent pour quantifier ces écarts, à savoir la somme des carrés desdits écarts ou le maximum des valeurs 35 absolues desdits écarts.

8

- on détermine comme langue du texte, la langue dont la distribution langue présente une plus grande similarité avec la distribution texte. Cette similarité est déterminée en ne retenant que les écarts qui souscrivent à un certain critère : présenter une valeur minimale. En variante, le critère auquel
5 doivent souscrire les écarts peut être de présenter une valeur inférieure à un seuil.

Le texte initial 15 est ensuite traité en fonction de cette détermination. Dans un exemple, il est traduit par le programme de traduction 13, au moyen d'un dictionnaire approprié 17 afin d'obtenir un texte final 19 écrit dans la
10 langue 18 choisie par un utilisateur de l'imprimante multi-fonctions 1 ; ledit texte final (s'affichant sur l'écran 4).

De nombreux autres traitements de documents écrits peuvent bénéficier d'une identification de la langue d'écriture. Par exemple :

- le document peut être classé dans la mémoire de données 7 de
15 façon différente en fonction de sa langue ;

- un programme de correction orthographique 14 peut sélectionner un dictionnaire adapté 17 plutôt qu'utiliser un dictionnaire général contenant toutes les langues possibles. Ils devraient donc gagner en vitesse, en consommation de mémoire et en qualité de résultat.

20 La figure 2 montre un exemple de comparaison d'une distribution de n-grammes d'un texte avec des distributions de n-grammes de trois langues distinctes α , β et γ .

Dans cet exemple de réalisation du procédé selon l'invention, on considère que le texte reçu est arbitrairement composé dans une langue
25 comprenant uniquement des lettres a, b c et d. L'alphabet de la langue est toutefois élargi d'un caractère spécial _ correspondant à l'espace présent au début et ou à la fin d'un mot. Le procédé de l'invention propose de fragmenter le texte en une pluralité de groupes de n caractères appelés n-grammes.

30 Avec des bigrammes, les lettres présentes dans la langue offrent un ensemble de 24 combinaisons de bigrammes possibles. Mais dans un texte, toutes les combinaisons ne sont pas forcément présentes. Ici par exemple, les bigrammes "aa", "a_" et "cc" ne sont pas écrits dans le texte. On établit alors une distribution texte de fréquences auxquelles apparaissent les
35 bigrammes dans le texte. Par distribution texte, on entend la statistique de

présence des bigrammes considérés dans le texte à étudier. Tout simplement, on compte le nombre d'apparitions de chacun d'eux dans le texte. On divise ce nombre par le nombre total de n-grammes, de bigrammes, du texte.

5 On a établi préalablement, sur un ensemble de textes connus dans une langue α , voire sur un dictionnaire de cette langue α , une distribution langue de fréquences auxquelles apparaissent les bigrammes dans cette langue α . On réitère cette opération pour un certain nombre de langues connues β et γ . Chaque langue connue β et γ est caractérisée par une
10 distribution de fréquences d'apparition de bigrammes. Ces distributions langues représentent la probabilité de présence des n-grammes, des bigrammes, dans la langue. Alors que la distribution texte est calculée par l'appareil 20, les distributions langues sont mémorisées à l'avance dans la mémoire 7.

15 Pour simplifier l'explication, on a représenté graphiquement ces distributions texte et langues α , β et γ . Sur cette représentation on a affiché les écarts calculés entre l'apparition d'un bigramme dans la langue et dans le texte. Ces écarts sont cotés en positifs ou en négatifs.

On compare ensuite lesdites distributions texte et langues α , β et γ
20 entre elles. Dans cet exemple de réalisation du procédé selon l'invention, la comparaison est réalisée en calculant la somme de toutes les valeurs absolues des écarts les séparant. Avec les valeurs affichées, le résultat de cette somme est 243 pour la distribution langue α , 451 pour la distribution langue β et 763 pour la distribution langue γ . La langue α présentant le moins
25 grand écart est donc choisie.

D'autres méthodes mathématiques existent pour quantifier ces écarts, à savoir la somme du carré desdits écarts. Dans ce cas, les résultats sont respectivement 4080 pour la distribution langue α , 11033 pour la distribution langue β et 29345 pour la distribution langue γ .

30 Une autre méthode est celle du maximum des valeurs absolues desdits écarts. Les résultats sont alors 27 pour la distribution langue α , 37 pour la distribution langue β et 73 pour la distribution langue γ .

Chacune des trois méthodes aboutit ici à la même conclusion : c'est la distribution langue α qui présente l'écart 243, 4080 ou 27 le plus faible par
35 rapport à la distribution texte. L'écart entre la distribution langue α et la

distribution texte, pour chaque bigramme, est symbolisé par la zone hachurée sur la figure 2,

La langue α présente une similarité maximale avec le texte ; elle est donc considérée comme étant celle du texte.

5 Il convient de mettre en évidence les particularités de la méthode liée au calcul des écarts. Par exemple, des fautes de frappe ou des mauvaises reconnaissances de caractères manuscrits pouvant avoir engendré ici le bigramme erroné "ac". En effet, ce dernier n'existe pas dans la langue α . D'une part l'écart de valeur 20 déterminé pour le bigramme "ac", entre la
10 distribution texte et la distribution langue α est négligeable dans l'opération d'addition décrite précédemment, au regard des vingt-trois autres bigrammes correctement distribués dans le texte.

D'autre part, il peut être envisageable de ne pas prendre en compte les n-grammes qui apparaissent dans le texte alors qu'ils n'apparaîtraient
15 pas dans la langue. Dans ce dernier cas, dans la comparaison, on ne retient les écarts que s'ils souscrivent à un critère. Le critère est ici celui de la présence du n-gramme dans la langue. Un autre critère est celui de négliger les écarts inférieurs à un seuil. Par exemple, si on néglige les écarts inférieurs à 10, la première méthode donne un résultat de 195 (au lieu de
20 243) pour la langue α , alors que le résultat est presque inchangé pour les langues β et γ . L'utilisation d'un critère permet ainsi d'accélérer le calcul et d'être plus sélectif. Une fois la langue du texte ainsi déterminée, le texte peut être traité. Dans ce dernier cas, il peut soit être édité tel quel, soit être corrigé, soit encore être traduit. Par exemple, si un texte est reçu dans une
25 langue α et si un utilisateur de l'appareil, ici du terminal mobile, a choisi une langue β en zone 18 dans la mémoire de données 7, on peut prévoir de traduire le texte de la langue α vers la langue β .

REVENDEICATIONS

1 - Procédé d'édition (12) d'un texte (15) exprimé dans une langue, comportant des étapes dans lesquelles :

- 5 - on reçoit, dans un appareil (1), une suite d'informations binaires correspondant à des caractères d'un alphabet ;
- on édite, sur un périphérique (4) de l'appareil, un texte correspondant aux caractères reçus ;
- on fragmente le texte en une pluralité de groupes de n caractères
- 10 appelés n-grammes, n étant un nombre entier positif non nul ;
- on établit une distribution texte de fréquences auxquelles apparaissent les n-grammes dans le texte ;
- on établit une distribution langue de fréquences auxquelles apparaissent les n-grammes dans une langue (α ; β ; γ) ;
- 15 - on réitère cette dernière opération pour un certain nombre de langues connues, chaque langue connue étant caractérisée par une distribution de fréquences d'apparition de n-grammes ;
- on compare lesdites distributions texte et langues entre elles ;
- on détermine comme langue du texte, la langue dont la distribution
- 20 langue présente une plus grande similarité avec la distribution texte et on traite le texte en fonction de cette détermination
- caractérisé en ce qu'il comporte l'étapes suivante :
- on calcule, n-gramme par n-gramme, l'écart entre la fréquence d'apparition d'un n-gramme dans le texte et dans la langue.
- 25 2 - Procédé selon la revendication 1, caractérisé en ce qu'on élargit les combinaisons possibles réalisées avec l'ensemble des lettres de l'alphabet de la langue avec un ou deux caractères d'espace () placés avant ou après un n-1gramme ou avant et après un n-2gramme.
- 3 - Procédé selon l'une au moins des revendications 1 à 2, caractérisé
- 30 en ce que les n-grammes sont des groupes de deux caractères appelés bigrammes.
- 4 - Procédé selon l'une au moins des revendications 1 à 2, caractérisé en ce que les n-grammes sont des groupes de trois caractères appelés trigrammes.
- 35 5 - Procédé selon l'une au moins des revendications 1 à 4, caractérisé

12

en ce que la comparaison des distributions comporte une étape d'addition, pour tous les n-grammes, de toutes les valeurs absolues des écarts les séparant.

- 5 6 - Procédé selon l'une des revendications 1 à 4, caractérisé en ce que la comparaison des distributions comporte une étape d'addition, pour tous les n-grammes, de tous les carrés des écarts les séparant.

7 - Procédé selon l'une des revendications 1 à 4, caractérisé en ce que la comparaison des distributions comporte une étape de détermination du maximum de toutes les valeurs absolues des écarts les séparant.

- 10 8 - Procédé selon l'une des revendications 1 à 7, caractérisé en ce qu'on ne retient les écarts que s'ils souscrivent à un certain critère.

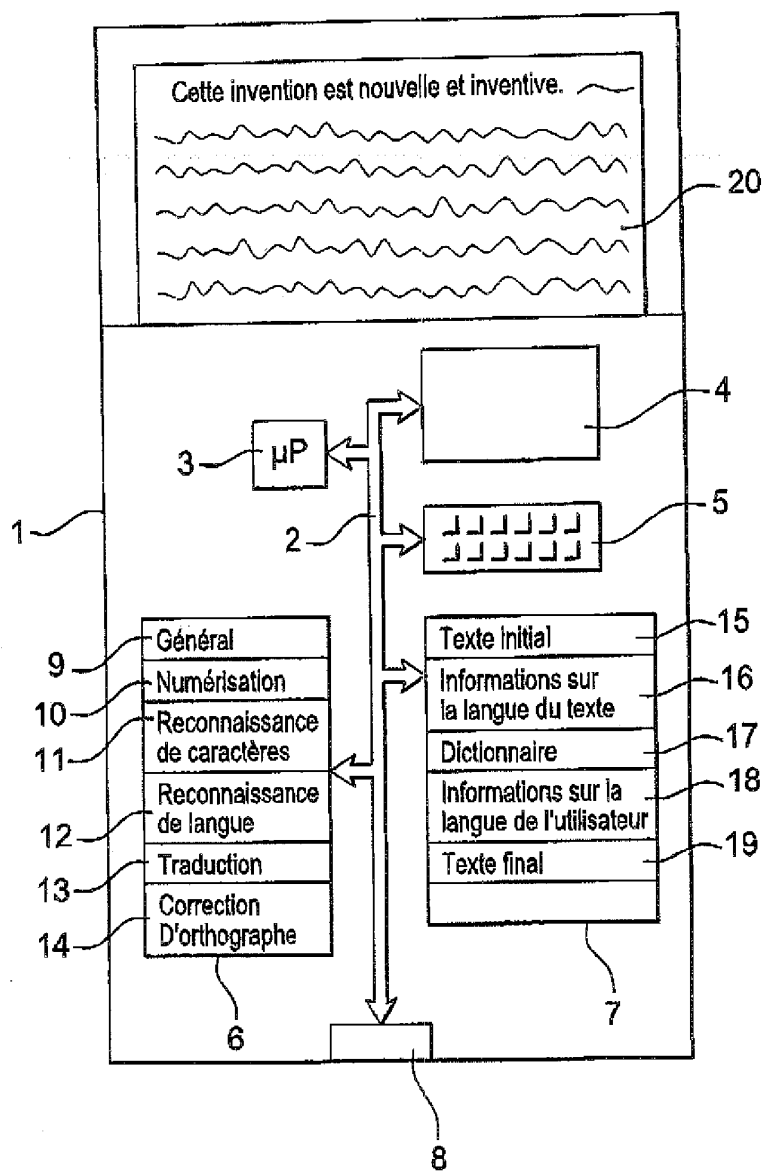
9 - Procédé selon l'une des revendications 1 à 8, caractérisé en ce que le critère auquel doivent souscrire les écarts est de présenter une valeur inférieure à un seuil.

- 15 9 - Procédé selon l'une au moins des revendications 1 à 8, caractérisé en ce que le critère auquel doivent souscrire les écarts est de présenter une valeur minimale.

- 20 10 - Procédé selon l'une au moins des revendications 1 à 9, caractérisé en ce que le traitement réalisé par le procédé est une correction d'orthographe (14).

11 - Procédé selon l'une au moins des revendications 1 à 9, caractérisé en ce que le traitement réalisé par le procédé est une traduction automatique (13) du texte reçu, dans une langue (18) paramétrée par l'appareil.

1/2

**Fig. 1**

2/2

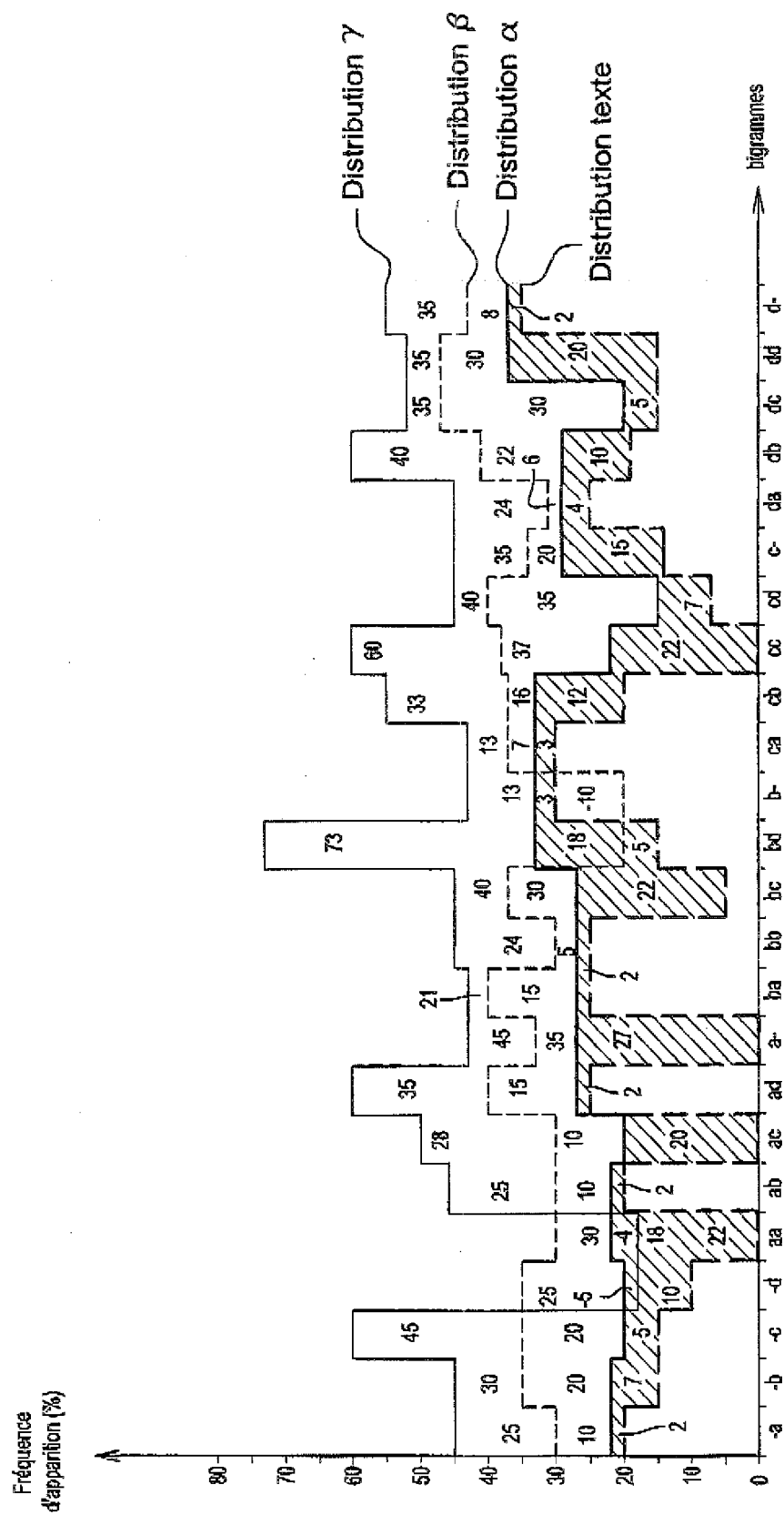


Fig. 2



RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications
déposées avant le commencement de la recherche

N° d'enregistrement
national

FA 692116
FR 0752564

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
X	CAVNAR W B ET AL: "N-GRAM-BASED TEXT CATEGORIZATION" 1998, PROCEEDINGS. ANNUAL SYMPOSIUM ON DOCUMENT ANALYSIS & INFORMATION RETRIEVAL, PAGE(S) 1-14 , XP002929596 * abrégé * * séquences 2,3, 4.1, 6 *	1-11	G06F17/27 G06F17/28 G06F17/18
X	SIMON KRANIG: "Evaluation of Language Identification Methods" [Online] 2005, , BACHELOR THESIS, UNIVERSITY OF TÜBINGEN , XP002451475 Extrait de l'Internet: URL:http://www.sfs.uni-tuebingen.de/iscl/T heses/kranig.pdf> [extrait le 2007-09-18] * abrégé * * séquence 4 *	1-11	
X	US 6 272 456 B1 (DE CAMPOS MIGUEL CARDOSO [US]) 7 août 2001 (2001-08-07) * colonne 2, ligne 50-56 * * colonne 3, ligne 4-14 * * colonne 4, ligne 37-65 * * colonne 5, ligne 15-32 * ----- -/--	1-11	
			DOMAINES TECHNIQUES RECHERCHÉS (IPC) G06F
		Date d'achèvement de la recherche 20 septembre 2007	Examineur Lopez, Patrice
CATÉGORIE DES DOCUMENTS CITÉS X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : arrière-plan technologique O : divulgation non-écrite P : document intercalaire T : théorie ou principe à la base de l'invention E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure. D : cité dans la demande L : cité pour d'autres raisons & : membre de la même famille, document correspondant			



RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications
déposées avant le commencement de la recherche

N° d'enregistrement
national

FA 692116
FR 0752564

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
X	PENELOPE SIBUN, JEFFREY C. REYNAR: "Language Identification: Examining the Issues" [Online] 1996, , 5TH SYMPOSIUM ON DOCUMENT ANALYSIS AND INFORMATION RETRIEVAL, LAS VEGAS, USA , XP002451476 Extrait de l'Internet: URL:http://citeseer.ist.psu.edu/rd/0%2C139 415%2C1%2C0.25%2CDownload/http://citeseer. ist.psu.edu/cache/papers/cs/810/http:zSzzS zwww.cis.upenn.edu/zS%7EjcreynarzSzsdair96 .pdf/sibun96language.pdf> [extrait le 2007-09-18] * abrégé * * séquences 1,3 * -----	1-11	DOMAINES TECHNIQUES RECHERCHÉS (IPC)
Date d'achèvement de la recherche		Examineur	
20 septembre 2007		Lopez, Patrice	
CATÉGORIE DES DOCUMENTS CITÉS X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : arrière-plan technologique O : divulgation non-écrite P : document intercalaire		T : théorie ou principe à la base de l'invention E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure. D : cité dans la demande L : cité pour d'autres raisons & : membre de la même famille, document correspondant	

Les renseignements fournis sont donnés à titre indicatif et n'engagent pas la responsabilité de l'Office européen des brevets, ni de l'Administration française

EPO FORM P0465