

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
15 January 2004 (15.01.2004)

PCT

(10) International Publication Number
WO 2004/006122 A2

(51) International Patent Classification⁷: **G06F 17/20**

(21) International Application Number:
PCT/CA2003/001023

(22) International Filing Date: 3 July 2003 (03.07.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/393,903 3 July 2002 (03.07.2002) US
10/289,656 7 November 2002 (07.11.2002) US
2,411,227 7 November 2002 (07.11.2002) CA

(71) Applicant (for all designated States except US): **2012244
ONTARIO INC.** [CA/CA]; 295 Phillip Street, Waterloo,
Ontario N2L 3W8 (CA).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **FUX, Vadim**

[IL/CA]; 575 Rustic Drive, Waterloo, Ontario N2K 2A4
(CA). **ELIZAROV, Michael, G.** [IL/CA]; 465 Kingscourt
Drive, Unit 3, Waterloo, Ontario N2K 3R5 (CA).

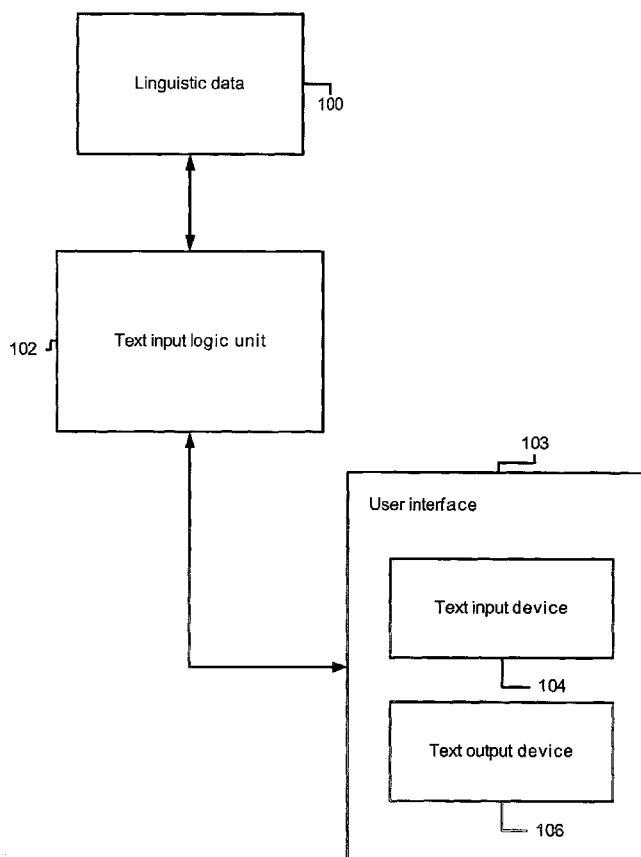
(74) Agents: **PATHIYAL, Krishna, K.** et al.; Research In Mo-
tion Limited, 295 Phillip Street, Waterloo, Ontario N2L
3W8 (CA).

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

(84) Designated States (*regional*): ARIPO patent (GH, GM,
KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW),
Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),
European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,
ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO,

[Continued on next page]

(54) Title: SYSTEM AND METHOD OF CREATING AND USING COMPACT LINGUISTIC DATA



(57) Abstract: A system and method of creating and using compact linguistic data are provided. Frequencies of words appearing in a corpus are calculated. Each unique character in the words is mapped to a character index, and characters in the words are replaced with the character indexes. Sequences of characters are mapped to substitution indexes, and the sequences of characters in the words are replaced with the substitution indexes. The words are grouped by common prefixes, and each prefix is mapped to location information for the group of words which start with the prefix.

WO 2004/006122 A2



SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW, ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY,*

BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW, ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)

- *of inventorship (Rule 4.17(iv)) for US only*

Published:

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

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