# The way we write

## Country-specific variations of the English language in the biomedical literature

*Rebecca Netzel, Carolina Perez-Iratxeta, Peer Bork & Miguel A. Andrade*

Ever since the 1950s, when research and engineering in the USA and, to a lesser extent, in the UK, started to expand dramatically, English has been the *lingua franca* of the scientific community (Garfield, 1998). As a consequence, many scientists all over the world are now obliged to describe their research and discuss their results in a language that is not their mother tongue. This clearly affects the communication of science in the worldwide academic community, because the way a researcher writes in English depends largely on his or her familiarity with the language.

For the sake of communicating science, the scientific community has to allow certain unavoidable differences in style, provided they are within the bounds of English grammar. But a scientist is not expected to be either a professional writer or a translator. Furthermore, there is no standard scientific English against which to compare a text, so it is difficult to evaluate the style of a scientific publication. In fact, there is not even a standard for the English language itself, as various countries, such as Canada, the Caribbean, India, the Philippines, New Zealand and the USA, have developed varieties of English that are as distinct from British English as they are from each other (Ritter, 2002).

> **...there is no standard scientific English against which to compare a text, so it is difficult to evaluate the style of a scientific publication**

Although it is not possible to define a common standard for written English in scientific communication, it is valuable to identify local peculiarities and differences in writing from authors from various countries. These clearly prevail in some journals more than others, depending on the level of copy-editing of the final text by editors and publishers. Such variations in the use of English, due to the authors' native language and cultural background, can not only make a text more difficult to understand and distract the reader from the content, but also hold the danger that the meaning and content of a sentence is diluted or misinterpreted by a reader with another language background. Thus, locally favoured words and phrases should be recognized, and eventually avoided, to increase the clarity of scientific communication.

To determine such variations in the scientific literature, we examined the MEDLINE database of biomedical articles (www3.ncbi.nlm.nih.gov/Entrez/index.html). This database contains more than 11 million references to biomedical articles, including the address of the main author, the country of the publisher and often an abstract of the publication. To associate abstracts with nationalities, we first extracted the name of the country from the affiliation field (Perez-Iratxeta & Andrade, 2002). We eventually restricted the study to the 50 countries with the greatest numbers of abstracts in the MEDLINE database (Table 1). Almost half of the publications selected were from a country where English is not the official language or where less than 10% of the population speak English as their first language. The grammatical analysis of the text was performed using the program TreeTagger, which is freely available software developed at the University of Stuttgart, Germany (www.ims.unistuttgart.de/projekte/corplex/TreeTagger), that associates a part-of-speech tag to each word in a text (see sidebars on page 448). We chose several parameters to illustrate the language variation observed for different countries.

First, we computed the average number of words and verbs per sentence and found that although these parameters vary greatly between countries, there are some correlations with the native language of a country (Fig. 1A). Anglo-Saxon scientists write longer sentences—an average of 27 words and 3.8 verbs per sentence for the UK—as would be expected from their familiarity with English. Another remarkable difference is seen in the implied involvement of the author in his or her research. This personal involvement can

**Table 1** | The 50 countries used in this study

| Country | *N* | Main languages | |
|---|---|---|---|
| Argentina | 12,450 | Spanish | |
| Australia | 83,531 | English | |
| Austria | 28,355 | German | |
| Bangladesh | 989 | Bengali | |
| Belgium | 41,411 | French | Dutch, German |
| Brazil | 28,152 | Portuguese | |
| Bulgaria | 3,370 | Bulgarian | |
| Canada | 145,630 | French | English |
| China | 42,511 | Chinese | |
| Czechoslovakia | 19,848 | Czech, Slovak | |
| Denmark | 39,018 | Danish | |
| Egypt | 5,163 | Arabic | |
| Finland | 34,783 | Finnish | |
| France | 186,848 | French | |
| Germany | 239,204 | German | |
| Greece | 14,903 | Greek | |
| Hong Kong | 6,119 | Chinese | English |
| Hungary | 14,149 | Hungarian | |
| India | 40,775 | Several | English |
| Ireland | 15,357 | English | |
| Israel | 39,437 | Hebrew | English |
| Italy | 151,889 | Italian | |
| Japan | 335,239 | Japanese | |
| Korea (South) | 16,977 | Korean | |
| Lebanon | 1,140 | Arabic | |
| Malaysia | 2,912 | Malay | |
| Mexico | 9,434 | Spanish | |
| Netherlands | 91,155 | Dutch | |
| New Zealand | 13,627 | English | |
| Nigeria | 5,184 | Niger-Congo | |
| | | Afro-Asiatic, Chadic | |
| Norway | 24,092 | Norwegian | |
| Pakistan | 2,102 | Several | English |
| Poland | 34,175 | Polish | |
| Portugal | 5,829 | Portuguese | |
| Puerto Rico | 1,075 | Spanish | |
| Romania | 2,919 | Romanian | |
| Russia | 28,030 | Russian | |
| Saudi Arabia | 5,547 | Arabic | |
| Singapore | 6,874 | Chinese | English |
| South Africa | 13,752 | Afrikaans, English | |
| Spain | 78,253 | Spanish, Others | |
| Sweden | 79,300 | Swedish | |
| Switzerland | 54,643 | French | German |
| Taiwan | 18,517 | Chinese | |
| Thailand | 6,375 | Thai | |
| Turkey | 16,213 | Turkish | |
| UK | 313,832 | English | |
| USA | 1,378,276 | English | |
| Venezuela | 4,123 | Spanish | |
| Yugoslavia | 11,395 | Several | |

In the period covered by our analysis (1991–2000), several countries either reunited or divided. For simplicity, we merged data from countries that belonged to the former republics of Yugoslavia, Czeckoslovakia, USSR and East and West Germany. *N*, number of abstracts analysed. Languages are assigned colours as in the key of Fig. 1.

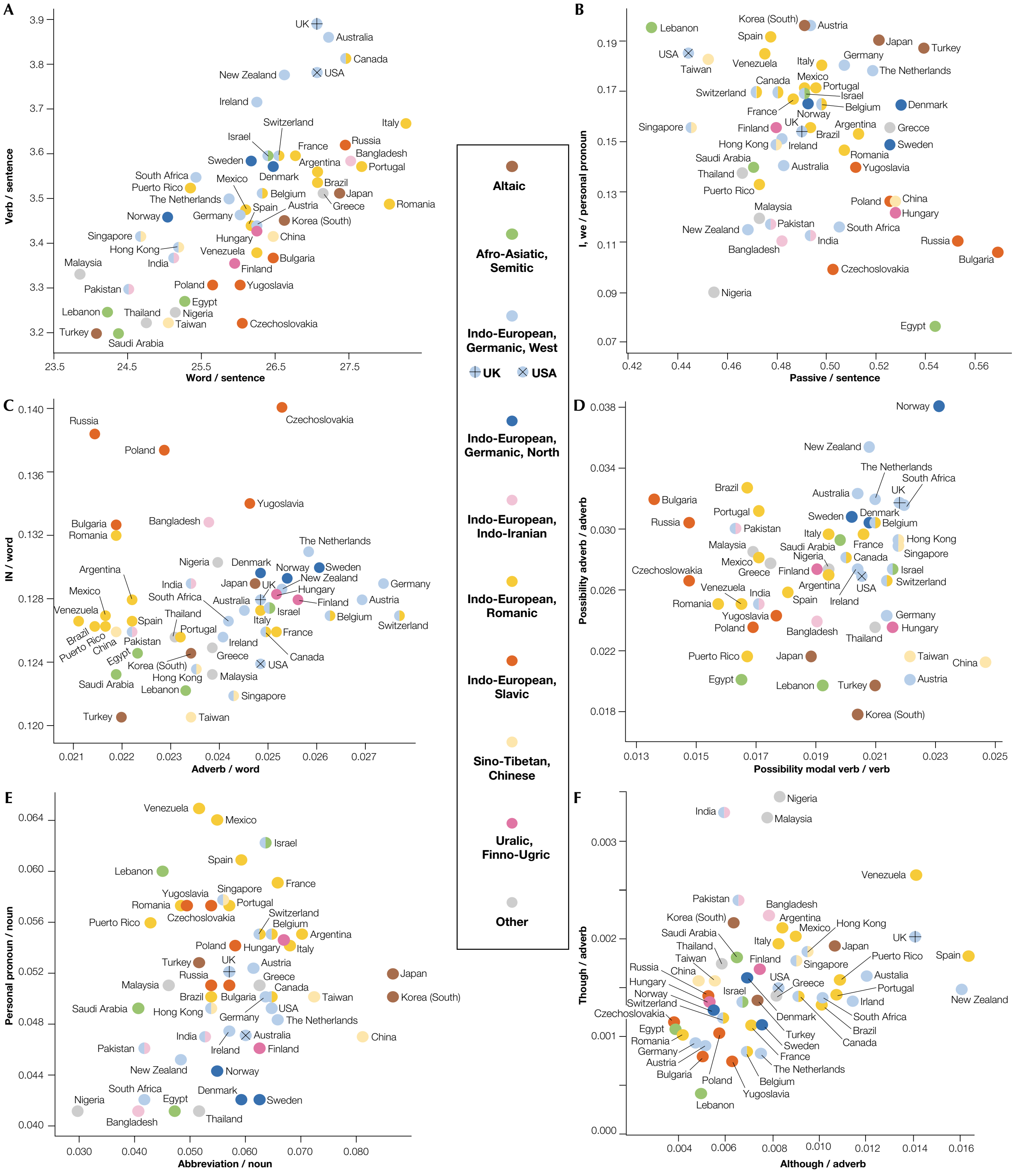

**Fig. 1** | Grammatical parameters by country. The colours indicate the language family of the main languages spoken in a country; the UK is marked with a plus sign and the USA with a cross for reference. Current data about the languages (official or not) that are spoken by at least 10% of the population in those countries, and data about the language families, were taken from the Ethnologue Database (www.ethnologue.com) and Beekes (1995). (**A**) The number of verbs per sentence versus the number of words per sentence. (**B**) The fraction of personal pronouns in the first person versus the number of passive constructions per sentence. (**C**) The fraction of all tags that are a preposition or subordinating conjunction (IN) versus the fraction of all tags that are an adverb. (**D**) The fraction of adverbs indicating possibility versus the fraction of modal verbs indicating possibility. (**E**) The number of personal pronouns per noun versus the number of abbreviations per noun. (**F**) The fraction of all adverbs that are 'although' versus the fraction of all adverbs that are 'though'.

be diminished by the use of the passive voice, which is discouraged in writing in general (Strunk & White, 1979), and in particular for technical writing (Day, 1994; Brown, 2000), but which nevertheless often pervades scientific articles (Möhn & Pelka, 1984). We distinguished passive sentences as those containing any form of 'be' followed by a verb in the past participle, allowing one adverb in between, such as "were significantly associated". Another indicator of personal involvement is the use of the first-person pronouns 'I' and 'we'. Fig. 1B plots these parameters, and shows a significant difference between the USA and the UK, with the USA standing out from the bulk of the Germanic countries in the top-left corner. Writers from Slavic countries occupy the opposite corner. Such an effect might also be related to the different role of the passive voice in some languages, for example Japanese and Russian, compared with English.

## USE OF TREETAGGER I

Tagging of a sentence extracted from MEDLINE entry PMID:10761406. The TreeTagger program annotates the words of a text with their part-of-speech tag—noun, adjective, verb, and so on—using word context (for more information, see Santorini, 1990). It identifies the words (left column), and assigns a tag (middle column) and the corresponding stem if the word is present in a lexicon (right column). Even if the word is absent from the lexicon (for example, 'capsular'), the tag is derived from the context of the word. See sidebar II for the definitions.

| Word | Tag | Stem |
|---|---|---|
| Two | CD | Two |
| cases | NNS | case |
| of | IN | of |
| late | JJ | late |
| postoperative | JJ | postoperative |
| capsular | JJ | <unknown> |
| block | NN | block |
| syndrome | NN | syndrome |
| that | WDT | that |
| occurred | VBD | occur |
| 4 | CD | 4 |
| and | CC | and |
| 8.5 | CD | cardinal number |
| years | NNS | year |
| , | , | , |
| respectively | RB | respectively |
| , | , | , |
| were | VBD | be |
| encountered | VBN | encountered |
| . | SENT | . |

The use of prepositions and adverbs also differs according to the local language (Fig. 1C). Writers from German-speaking countries, for instance, use many adverbs compared with Spanish speakers; indeed, the two languages differ considerably in the way they form adverbs and use them in a sentence. An example is the expression "sorfältig statistisch ausgewertet" in German, meaning "carefully statistically evaluated". The literal Spanish version "cuidadosamente estadísticamente evaluado" sounds odd, and Spanish speakers would rather write "evaluado con un método estadístico de manera cuidadosa", which literally translates to "evaluated with a statistical method in a careful way". This substitutes the adverbs with equivalent noun–adjective pairs. Scientists from Slavic countries stand out as using many prepositions, which is in contrast to writers from several Asian countries.

Scientific language should be clear, conclusive and unequivocal. However, scientists often use words that imply uncertainty, such as the modal verbs 'would', 'could', 'should', 'may' or 'might', or adverbs such as 'likely', 'possibly' or 'probably'. Anglo-Saxon countries are prominent in this respect (Fig. 1D), whereas Chinese, Altaic and German-speaking countries tend to avoid such adverbs and modal verbs. There is also a country-specific difference in the use of nouns (Fig. 1E). These words can be substituted by a personal pronoun (for example, "It was isolated from kidney"), referring to the use of the noun in a previous sentence (such as "Protein X has a low molecular weight.") This back-referring is more common among authors from Romanic countries, particularly those that are Spanish-speaking, who use the most personal pronouns per total number of nouns in a text. Another common way to substitute nouns is by abbreviation, which is more prevalent among scientists from those Asian countries with ideographic writing, who tend to formulate shorter representations of many words.

Another good marker for local peculiarities are words that can be used interchangeably. In our analysis, we chose the pairs 'may/might' and 'though/although' (Fig. 1F). Papers by Anglo-Saxon writers show the highest prevalence of 'although' and 'may'. By contrast, scientists from India are fond of using 'though', which is another example

## USE OF TREETAGGER II

A corpus of 3,754,882 abstracts from MEDLINE, which represents 50 countries, was tagged in 26 h using a 550 Hz PIII CPU. The number of tags produced was 774,936,102, which is an average of 207 tags per abstract.

| Tag | % use | Part of speech |
|---|---|---|
| ***Noun-related*** | | |
| NN | 20 | Noun, singular or mass |
| NNS | 8 | Noun, plural |
| NP | 4 | Proper noun, singular |
| NPS | <1 | Proper noun, plural |
| FW | <1 | Foreign word |
| DT | 8 | Determiner |
| PP | <1 | Personal pronoun |
| PP$ | <1 | Possessive pronoun |
| PDT | <1 | Predeterminer |
| POS | <1 | Possessive ending |
| WP | <1 | Wh-pronoun (relative pronoun) |
| WP$ | <1 | Possessive wh-pronoun |
| EX | <1 | Existential 'there' |
| JJ | 11 | Adjective |
| JJR | <1 | Adjective, comparative |
| JJS | <1 | Adjective, superlative |
| ***Verb-related*** | | |
| VB | 2 | Verb, base form |
| VBD | 3 | Verb, past tense |
| VBG | 1 | Verb, gerund or present participle |
| VBN | 3 | Verb, past participle |
| VBP | 1 | Verb, non-third-person singular |
| VBZ | 1 | Verb, third-person singular present |
| MD | <1 | Modal |
| TO | 2 | To |
| ***Adverbs*** | | |
| RB | 2 | Adverb |
| RBR | <1 | Adverb, comparative |
| RBS | <1 | Adverb, superlative |
| RP | <1 | Participle |
| WDT | <1 | Wh-determiner |
| WRB | <1 | Wh-adverb |
| ***Other*** | | |
| CC | 4 | Coordinating conjunction |
| CD | 4 | Cardinal number |
| IN | 12 | Preposition or subordinating conjunction |
| LS | <1 | List item marker |
| UH | <1 | Interjection |
| SYM | <1 | Symbol |
| ***Punctuation*** | | |
| SENT | 4 | Sentence |
| , | 4 | Comma |
| : | <1 | Colon |
| ( | 1 | Opening bracket |
| ) | 1 | Closing bracket |

**Table 2** | Most frequently used words in various countries

| Country | Adjectives | Nouns | Verbs | Adverbs | Example sentence | PMID ref |
|---|---|---|---|---|---|---|
| Spain | Infrequent, bibliographic | Repercussion, evolution, existence, **sunflower**, **olive**, **wine** | – | Basically | Prevalence of CYP2D6 gene duplication and its **repercussion** on the oxidative phenotype in a white population. | 7697944 |
| Japan | Useful | **Bullfrog**, **shadow** (in radiography) | Clarify | Faintly, next, suddenly, scarcely | MDR-1 protein was **faintly** expressed in one of four chemoresistant patients, but Bcl-2 were [**sic**] clearly detected in four patients. | 12538495 |
| UK | Unsuitable, unlinked, unfamiliar | Marmoset, **consultant**, **questionnaire** | Lie, mirror, arise, tackle | Wholly, principally, particularly | The morphology of these projection neurons was revealed in great detail and confirmed that the projection **arises wholly** from pyramidal cells. | 11602231 |
| Russia | **Gravitational** | (Space) **mission**, **quantum**, **hibernate**, peculiarity, regularity, realization | – | **Thermo-dynamically** | The article is devoted to the question of **peculiarity** of bronchopulmonary system's pathology in the workers of the animal fodder production [**sic**]. | 10341521 |
| India | **Malarial**, -wise (as in stepwise), ascorbic | Malaria, buffalo, peanut, garlic, catfish, | Impart (convey) | Appreciable | Hydroxypropylmethylcellulose (HPMC) was used to **impart** strength and sphericity to the agglomerates. | 12476867 |
| France | Exceptional, digestive | Trouble | Envisage (imagine) | Successively (sequentially), essentially, sometimes | These 2 cells [**sic**] lines being able to clone, it is hard to **envisage** clonogenic assays. | 3051563 |
| China | **Medicinal**, **radiant** (heat), **noxious** (heat) | **Acupuncture**, **coal**, **tea** | **Burn**, replenish, alleviate | Obviously, meanwhile | Because only a catalytic amount of ERK2/pTpY is required, this method **alleviates** the need for large quantities of phospho-ERK2. | 12056917 |
| Germany | Satisfying practicable, unremarkable | Hint, precondition multitude | – | Additionally, exactly, | In clinically presumed spontaneous spinal cord infarction and **unremarkable** signaling of the spinal cord during sequential MRI investigations vertebral body infarction may serve as the only confirmatory sign of spinal cord ischemic stroke. | 11987007 |
| US | **Federal**, investigational, supplemental | Residency, **cocaine**, **payment**, **veteran**, **reimbursement**, physician, care, plan, **noncompliance**, effort, **profit** | Sponsor, mandate | – | Loss of revenue, mainly from **noncompliance** with charge capture resulted in the hospital billing only US$386,794.32 with a total **reimbursement** of US$165,779.86. | 12488156 |

Words in bold typeface have specific meanings and are probably related to local research rather than to local language usage. The bold and underlined words in the example sentences indicate the most abundant country-specific terms. The words shown were found to be more common in the abstracts of the corresponding country than in the abstracts of any other of the 19 representative countries (as in Fig. 2). Note that most of the sentences are grammatically correct, but the usage of the marked (bold and underlined) words is unusual.PMID ref, PubMed reference number.

of how a country develops its own norms in the use of English. Finally, we analysed which words are specifically used in the scientific literature from these 50 countries (Table 2). Some of these words indicate a focus on certain research fields in a country, but others indicate language usage or even social differences between countries.

Clearly, this study has its limitations, as it takes raw data from MEDLINE abstracts that represent only the biomedical literature. Also, the authors' affiliations do not necessarily indicate the real distribution of a publication's authors, as exemplified by this article, which has been written by German and Spanish scientists from German institutions, communicating to each other in a kind of English. Nevertheless, there are detectable differences in the use of English in the publications from the countries that we analysed. The most obvious factor is, of course, the local language in a country, as indicated by the clustering of countries using the same language or languages of the same family in Fig. 1. But these groupings are not perfect, and there is a great variation in the use of English depending on the parameters used in the study.

In addition, other cultural and geographical factors have a role in the variability of scientific English. For instance, the mobility of the scientific community that puts scientists of different countries in contact is one such factor. In general, scientific communication, recently made easier by the worldwide web, email and electronic journals, could contribute to a convergence towards a global consensus for the English language. But such a consensus may not be 'proper' English as defined by a British or US dictionary. To illustrate this point, there are many words that are already more broadly accepted in the non-native English-speaking community (Fig. 2). We think that these words are preferred by non-native speakers because they are more simple and easier to interpret, whereas a native speaker would find these words either too colloquial or would choose a synonym from a wider range of words with more particular gradations of meaning. This situation is not limited to the international scientific community but takes place in other settings as well.

> **...scientific communication ... could contribute to a convergence towards a global consensus for the English language**

It is not yet clear whether this situation constitutes an impoverishment or an improvement of the English language. What is clear is that current atypical word usage by various countries can make communication more difficult. An example is the use of the term 'subvention', which is used for 'grant' or 'subsidy' in the Brussels administration of the European Union, but which is not a common term for native English speakers. Other examples are the German bastardized term 'handy' for a mobile phone or the product name 'Bitter Sin' in Spain, a drink that is bitter

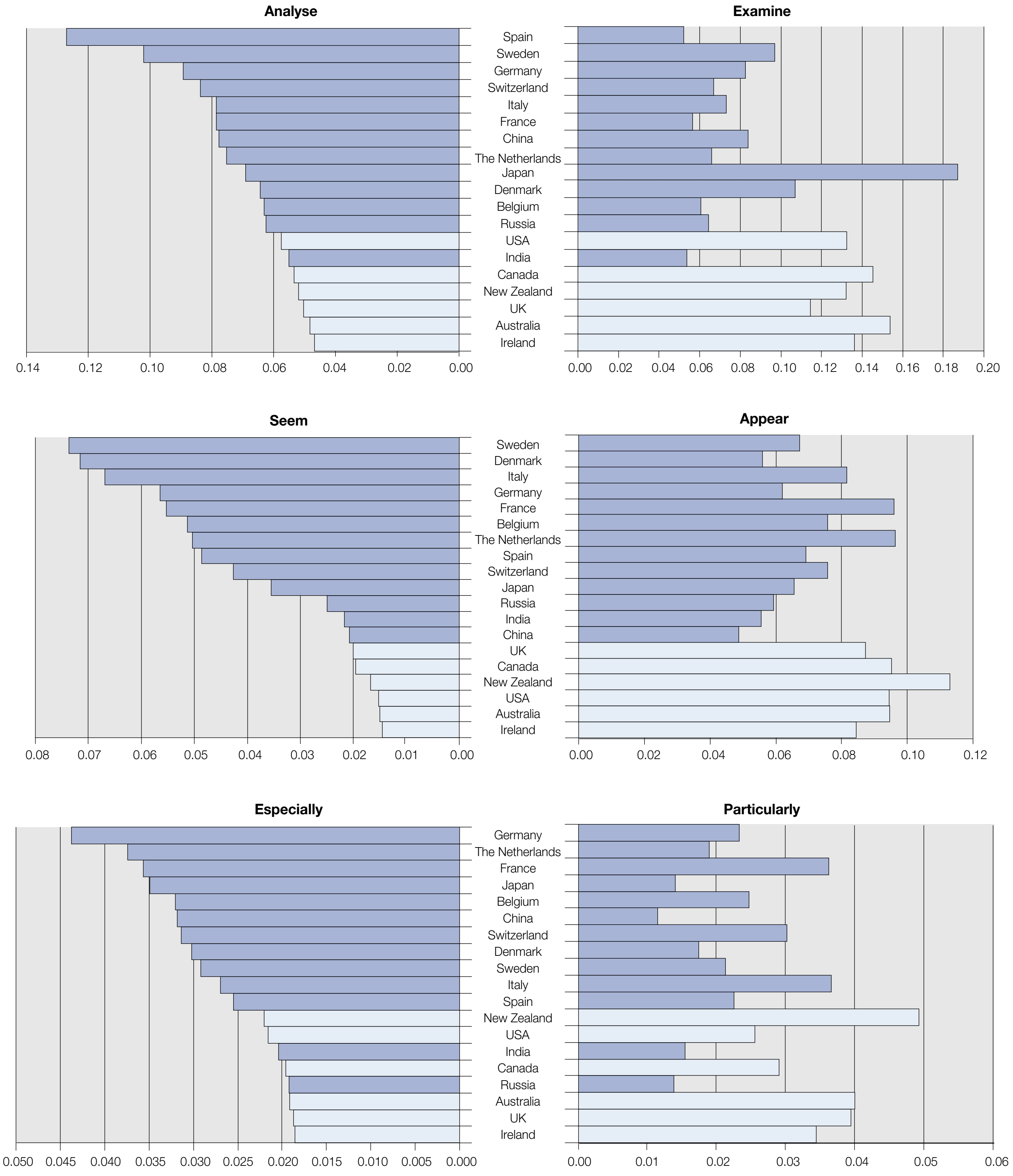

**Fig. 2** | Word usage per abstract for 13 countries in which English is not the main language (dark blue bars) and for 6 countries in which it is (light blue bars). Left-hand side, words not generally used in English-speaking countries. Right-hand side, equivalent words commonly used by native English speakers. The verb 'to analyse' can be interpreted in the sense of examine (inspect), study (as when learning), evaluate (weighing up), explore (discover) or dissect (cut up). 'Seem' and 'appear' can have equivalent meanings. However, 'seem' is preferred among non-native English speakers, as 'appear' has the associated alternative meaning of 'becoming visible', which 'seem' does not have. The adverb 'especially' could mean 'particularly' (principally) or 'exceptionally' (remarkably). In this respect, it could be considered by expert users of English as too vague, and 'particularly' might be preferred.

> …for the sake of clarity of communication, divergence in scientific writing should be minimized or at least slowed down…

and non-alcoholic—'sin alcohol'. But, as we pointed out earlier, there is no norm for the English language, so such developments are not necessarily bad, provided that they conform to syntax rules. Nevertheless, for the sake of clarity of communication, divergence in scientific writing should be minimized or at least slowed down, so that deleterious innovations can be recognized and weeded out, and scientists will be able to understand each other better.

To keep divergence at bay, teaching of the English language is probably not sufficient, as local teachers may further spread particular local biases and variations. Much more important is regular contact between scientists from various countries, particularly with native English speakers, which would help all concerned to adhere to a standard form of scientific English. This does not necessarily mean face-to-face communication, but could also occur through reading of scientific literature published in English. In this respect, the editorial control of published material has an increasingly important function. The evolution of scientific English as a variant form of English should be seen as a healthy development and may improve communication in due course. Human languages have changed over centuries, and English itself was enriched by both Roman and Norman invasions. We should therefore not fear for the English language when it is again invaded by hordes of scientists from all over the world, albeit much more peacefully.

ACKNOWLEDGEMENTS

We thank H. Schmid for providing TreeTagger and E. Minch for comments on the manuscript. We are grateful to the National Library of Medicine for licensing MEDLINE to us. We thank the members of our international group, who are expert users of many languages, for discussions: F. Ciccarelli (Italian); M. Suyama (Japanese); S. Schmidt, J. Korbel and C. von Mering (German); Y. Zdobnov (Russian); I. Letunic (Croatian); D. Torrents (Catalan, Spanish); W.C. Lathe III (US English, Korean); Y.P. Yuan (Chinese-Mandarin); P. Shah (Hindi, Gujarati, English); and D. Jaeggi (British English).

## REFERENCES

Beekes, R.S.P. (1995) *Comparative Indo-European Linguistics: An Introduction* (John Benjamins, Amsterdam, The Netherlands).

Brown, B.W. (2000) *Successful Technical Writing: A Practical Approach* (Goodheart-Willcox, Illinois, USA).

Day, R.A. (1994) *How to Write and Publish a Scientific Paper*, 4th edn (Cambridge Univ. Press, UK).

Garfield, E. (1998) Mapping the world of science. Paper presented at the 150th Anniversary Meeting of the American Association for the Advancement of Science, Philadelphia, Pennsylvania, USA, 14 February 1998. <www.garfield.library.upenn.edu/papers/mapsciworld.html>.

Möhn, D. & Pelka, R. (1984) *Fachsprachen—Eine Einführung* (Max Niemeyer, Tübingen, Germany).

Perez-Iratxeta, C. & Andrade, M.A. (2002) Worldwide scientific publishing activity. *Science*, **297**, 519.

Ritter, R.M. (2002) *The Oxford Guide to Style* (Oxford Univ. Press, UK).

Santorini, B. (1990) *Part-of-speech tagging guidelines for the Penn Treebank Project* (Technical Report MS-CIS-90-47, Department of Computer and Information Science, Univ. of Pennsylvania, USA).

Strunk, W. & White, E.B. (1979) *The Elements of Style*, 3rd edn (Allin & Bacon, Massachusetts, USA).

**Rebecca Netzel is at the Institute of Translation and Interpreting at the University of Heidelberg, Germany**

**Carolina Perez-Iratxeta is at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany.**

**Peer Bork is at the EMBL in Heidelberg, Germany.**

**Miguel A. Andrade is at the Ottawa Health Research Institute (OHRI), Canada.**
**E-mail: mandrade@ohri.ca**

doi:10.1038/sj.embor.embor833