Is English still the **lingua franca** of the state-of-the-art?

The world's most complete IP data collection

LIGHTHOUSE IP

Contents

Is English still the lingua franca of the state-of-the-art?

Much has been written and discussed in patent searching circles about the enormous numbers of Chinese-language publications which seem to be swamping the weekly updates of patent databases. However, the modern Chinese patent law has only been in force for some four decades, whereas the cumulated state-of-the-art from Western European and United States sources dates back to at least the late 19th or early 20th centuries.

All of these older documents are still theoretically relevant as novelty-destroying prior art, since our working definition of "new" places no limits on the age of a disclosure. Although some technical fields have a reasonably precise "date of birth", others do not; for the mechanical arts in particular, damaging citations can emerge from many years before the priority date of a new application, as old solutions are bypassed, forgotten and then rediscovered by a new generation of inventors. For both searchers and database producers, it may be useful to explore whether in fact the balance of publication language as a proportion of the entire state-of-the-art has already tipped in favour of Chinese or other Asian languages, rather than English.

Transitions between different languages of publication have been observed before. In the dawn of modern scientific method, many experts published their findings in the common language of international discourse, which could be Latin or French. ***Figure 1*** shows the drift away from these languages and towards English over the course of three centuries, for a journal series which has existed over the entire period. The same phenomenon holds true in specific disciplines; by the late 19th century, German was the dominant publication language in organic chemistry, led by the powerful dyestuffs industry in that country, and the natural language to use for the creation of edited compilations such as the multi-volume Beilstein and Gmelin "handbooks".

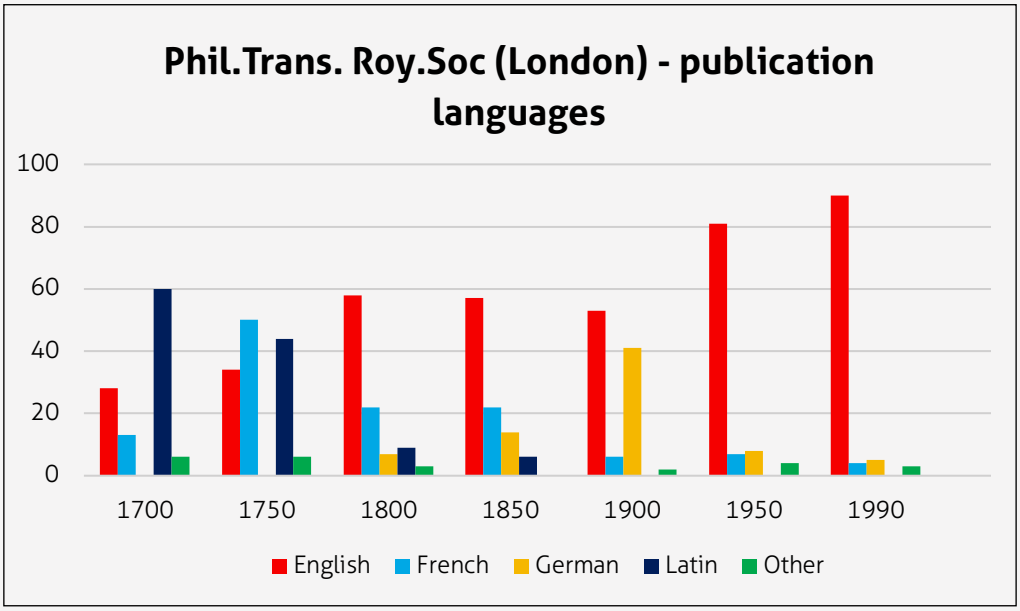

Figure 1: *publication languages of the philosophical transactions of the royal society*

Changes in international patent filing

Before considering the place of national publications in the patent sphere (applications, granted patents and utility models), it is worthwhile considering how the PCT system has changed the processes of international filing since it came into operation in 1978. From the beginning of the Treaty, the rules allowed for filing of applications in many languages, but restricted publication of the 18-month unexamined applications to one of only five alternatives; English, French, German, Japanese or Russian.

If an applicant filed their application in a language other than one of these five and did not additionally supply a translation into one of the publication languages, the International Search Authority would prepare a translation into English, which was used as the preferred language of publication. Over the following decades, additional languages were accepted as publication languages, starting with Spanish from 1985 and up to the current total of ten since 2009 (***see Table 1***; column headers are the ISO 639 language codes).

Period in effect	En	Fr	De	Jp	Ru	Es	Zh	Ar	Kr	Pt
1970.06.19-1984.12.31	Y (*)	Y	Y	Y	Y					
1985.01.01-1993.12.31	Y (*)	Y	Y	Y	Y	Y				
1994.01.01-1998.06.30	Y (*)	Y	Y	Y	Y	Y	Y			
1998.07.01-2002.12.31 (#)	Y	Y	Y	Y	Y	Y	Y			
2003.01.01-2006.03.31 (§)	Y	Y	Y	Y	Y	Y	Y			
2006.04.01-2008.12.31	Y	Y	Y	Y	Y	Y	Y	Y		
2009.01.01- date	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

(*) for applications as filed and all ISA-prepared English translations of applications (Rule 48.3(b)).
(#) Rule 48.3(a-bis) in force provided that an applicant-prepared translation if supplied under Rule 12.3 was used as publication text; otherwise, Rule 48.3(b) applied and ISA prepared an English translation.
(§) Rule 48.3(a-bis) deleted. After 2003.01.01, applicants are required to provide their own translation under Rule 12.3 or 12.4.

Table 1: *publication languages of the philosophical transactions of the royal society*

After the adoption of Chinese as a PCT publication language in 1994, the initial impact upon the languages appearing in any one weekly or yearly update was not very marked, but as time went on, there was increasing downward pressure upon English as part of the language mix. By the time that Korean was adopted in 2009, annual English-language PCT publication had dropped from over 70% to just over 60% of the total, and combined Asian languages (Japanese, Chinese and Korean) increased to just over 20%. Around 2016, English dropped below the 50% level for the first time, and increasing pressure had further reduced the proportion published in the other six languages, with Portuguese, Arabic and Russian barely making 1% combined. By 2023, combined Asian languages made up nearly 50% of the annual total of publications. ***Figure 2*** illustrates the overall trends, as a normalised proportional chart rather than absolute numbers of publications.

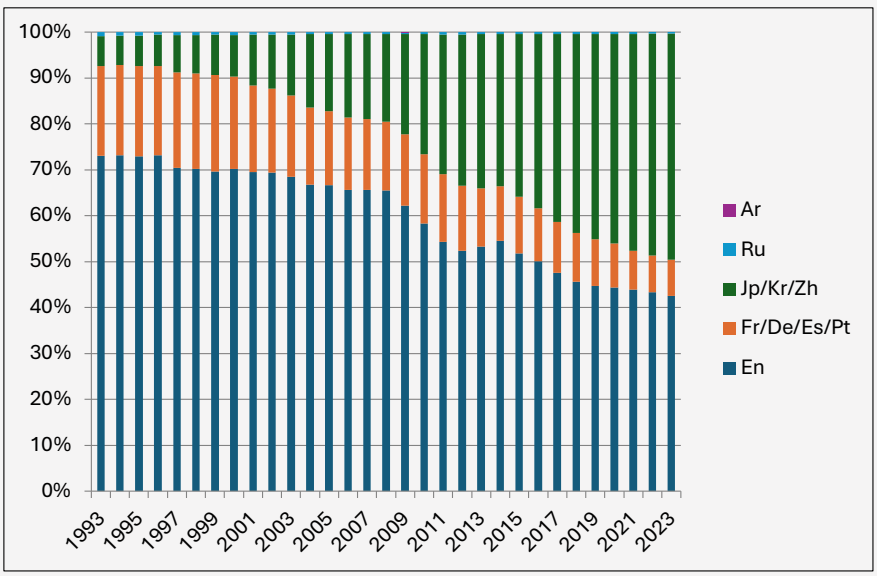

Figure 2: *Annual changes in language use under the PCT*

To a generation of searchers which had grown accustomed to the "problem" of large numbers of Japanese-only families through the 1970s-1980s, the advance of Chinese was startling. Further analysis of the Asian languages compared to English, also based on the annual proportion of publications, is shown in ***Figure 3***. It is clear that all three languages have advanced at the expense of English; since 2009 when

Korean was allowed as a publication language, its use has grown to 7% of output, Chinese has expanded hugely from less than 4% to over 24% and Japanese has maintained its position of around 17% of publications.

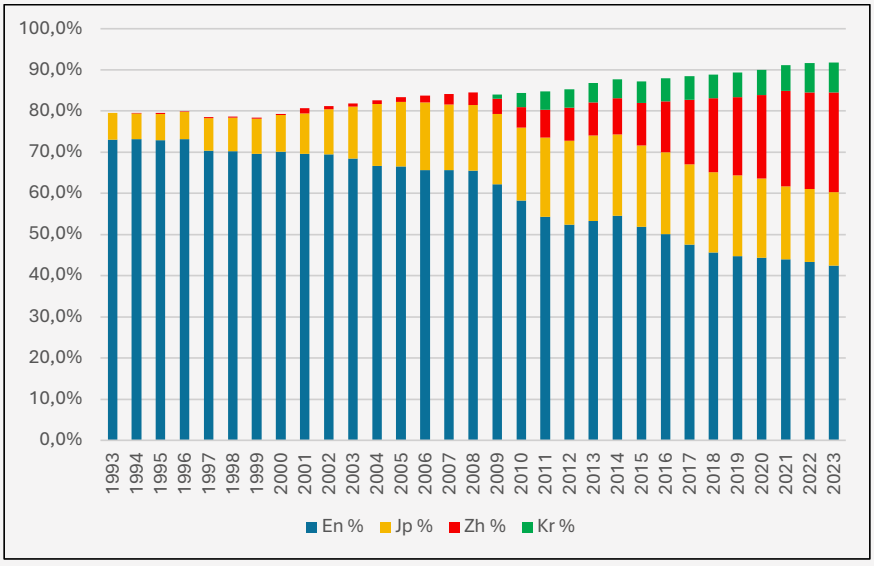

Figure 3: *English vs. Asian languages: annual proportion of PCT publications*

As the result of the rapid increase in the proportion of PCT publications in Asian languages, the patentability searcher would expect that the most recently-published prior art would also begin to be characterised by a rising proportion of Chinese, Japanese and Korean documents. However, as stated above, it is not always the most recent publications which are the best prior art, so we need to consider also the impact of Chinese upon the state-of-the-art as a whole. One estimate can be obtained using the same PCT data, but considering instead the absolute numbers of publications rather than the proportions in any one year, and cumulating the data each year to show the growth in total numbers of PCT publications entering the state-of-the-art.

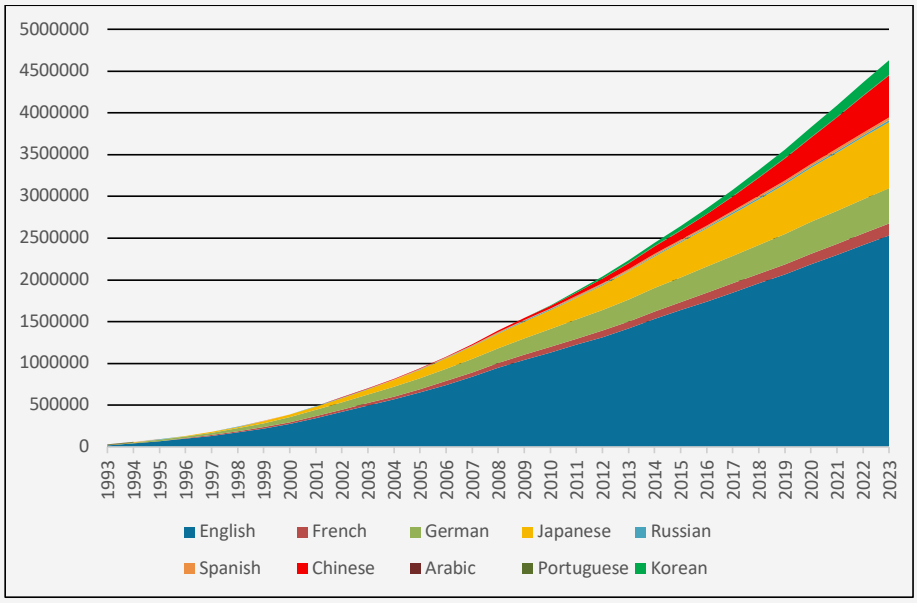

Figure 4: *Cumulative publication of WO-Adocuments, by language*

Figure 4 shows that the state-of-the-art – as represented by the cumulated PCT international applications published since 1993 – has clearly shown a progressive reduction in the prominence of English as a publication language. However, from the point of view of the prior art searcher, the issue is not the annual changes but whether this trend has impacted the overall proportion of English in the cumulated state-of-the-art i.e. the area under the graph.

To estimate these numbers, it is possible to take a number of snapshots of specific years and show how the cumulated totals have varied. Table 2 shows the year in which WO-A publications passed a significant milestone, and the percentage of the cumulated total which had been published in English, Japanese, Chinese and Korean at that point in time.

Year	Total WO-A publication exceeded	% English	% Japanese	% Chinese	% Korean
2002	500,000	70.6	8.7	0.5	0.0
2006	1,000,000	68.9	11.5	0.8	0.0
2012	2,000,000	64.5	14.5	2.7	1.1
2017	3,000,000	60.0	16.3	5.6	2.5
2021	4,000,000	56.2	17.0	9.2	3.4

Table 2: *Proportion of publications in the PCT collection, at selected milestone dates*

These data suggest that – all other things being equal – approximately half of the citations in a present-day search through the PCT collection only would consist of publications in English. Of course, this assumes that any documents relevant to the case being investigated are evenly distributed across all years and from all possible countries of origin, and that PCT published applications are the exclusive source of the state-of-the-art.

However, neither of the first two approximations is true; some countries are more prominent sources of new additions to the state-of-the-art than others (influencing the language of publication), and it is usually true that fewer relevant documents are found in earlier years compared to the most recent period prior to filing (i.e. there is a form of "literature relevance half-life" which skews the date of prior art citations).

As to the third assumption, it is clear that PCT published applications are not the only patent documents recognised as being in the state-of-the-art; a modern patentability search must include consideration of national collections of documents as well. The next step in our analysis must consider the impact of Asian languages upon the global state-of-the-art.

The impact of national document collections

The PCT Regulations have tried to address the issue of the scope of a comprehensive search by establishing a minimum baseline document collection for the work of the International Search Authorities (ISAs). Rule 34 defines this so-called Minimum Documentation, which the ISAs are mandated to search under Article 15(4) of the Treaty, which states:

"The International Searching Authority ... shall endeavor to discover as much of the relevant prior art as its facilities permit, and shall, in any case, consult the documentation specified in [Rule 34]."

The corresponding text of Rule 34 has been modified during the operational life of the Treaty. ***Table 3*** indicates how the scope of the patent literature as defined in the Minimum Documentation definition has evolved over time, excluding consideration of the non-patent literature as defined in Rule 34.1(b)(iii). It is clear that the definition has been adapted as the perceived significance of patent publications in Asian languages has increased. We may gain some further insight by considering the contribution made by selected national data sets which are published in English (notably the United States, United Kingdom and some of the EPO output), Japanese, Chinese or Korean.

Period	National patents from 1920 onwards (*)	International and regional systems
1970-1984	FR-A, -B, -U DE-A, -B JP-A, -B SU-A1, -A2, -A3, -A4 CH-B GB-A, -B US-A, -B + additional En, Fr, De basics without priority	WO-A 1978+ EP-A, -B 1978+ OA-B, -U 1982+ AP-B 1984+
1985-1998	as above + additional Es basics without priority	as above + EA-A, -B 1996+
1998-2007	as above + RU-B	
2007-2012	as above + KR-B	
2012-date	as above + CN-B	

(*) For the sake of simplicity, the KD codes -A, -B and -U are used as a shorthand to represent unexamined published applications, granted patents and published utility model applications and/or registrations respectively, irrespective of the actual codes used by the patent office in question. Where specific series of documents did not commence publication until after 1920, the collection is included from the earliest date available.

Table 3: *Abridged definition of the PCT Minimum Documentation since 1970*

When analysing multiple documentation sets, it is hard to avoid a degree of double-counting. For example, some national publications will be in the form of a national-language translation of a previously-published PCT application; they add to the number of documents in the state-of-the-art without adding new information per se.

Some national published applications will be re-published later as a granted patent (and be counted twice), whilst others will only enter the state-of-the-art as a single unexamined application and never mature into a grant. All the national grants arising from a single priority application (PCT or otherwise) will add to the numbers of documents within the cumulated state-of-the-art, even though they add little new information about the essential invention.

Some of these issues could be resolved by analyses of patent family structures (e.g. by filtering to include only 'single-member' families or the first-published member of a national family) or using information embedded in publication/application number formats. If the authority concerned uses the current

WIPO ST.13 standard including the "type of IP right" two-digit prefix to differentiate between national applications filed directly or via the PCT route, this could be used to filter out duplicate publications. Similarly, some authorities (such as Japan) use a different publication number series to denote translations into the national language of a previous PCT document.

However, for the purposes of this analysis no such filtering has been applied, and the following data represent a simplistic attempt to quantify the impact of Chinese documentation upon the language balance in the overall state-of-the-art. Although the numbers of national documents could be derived more accurately by using the national authority files deposited at WIPO under st.37, these files are predominantly presented in a form which cumulates across all publication years and are somewhat difficult to use to extract a single year's output. The statistics for national publication were therefore derived from searches on the DPMA Depatisnet Service, WIPO's Patent-scope and the EPO Espacenet Worldwide database, and shown in ***Table 4.***

Year	2002	2006	2012	2017	2021
Publications in English					
GB-A, -B	13564+8691	11846+7907	10653+6864	11768+6310	11267+10858
US-A1	198557	293830	330678	372081	409377
US-B1, -B2	140020+27391	27670+146111	17981+235180	29182+289655	29128+298297
EP-A1, -A2 (calculated as 60% of total)	38152+23503	50141+25851	60511+22173	85298+5271	99176+3892
EP-B1 (calculated as 60% of total)	28435	37671	39386	63389	65291
Sub-total (national/regional English)	**478312**	**601027**	**723426**	**862954**	**927286**
Publications in Japanese					
JP-A	419118	382181	291442	268439	234631
JP-B1, -B2	463+117319	1050+133147	4093+258004	10909+188683	11297+170071
JP-U, -Y2	8122+200	11305+3	8130+0	6028+0	5498+0
Sub-total (national Japanese)	**545222**	**527686**	**561669**	**474059**	**421497**
Publications in Chinese					
CN-A, -B	45324+207	110770+21543	334063+229471	832441+510486	1009021+752348
CN-U, -Y	0+59927	0+103386	539762+0	973293+0	3119990+0
Sub-total (national Chinese)	**105458**	**235699**	**1103296**	**2316220**	**4881359**
Publications in Korean					
KR-A, -B1	97488+45810	135597+120994	139465+113182	143456+119540	156393+146596
KR-U, -Y1	22+40290	201+30017	9008+6333	4425+2977	2916+1835
Sub-total (national Korean)	**183610**	**286809**	**267988**	**270398**	**307740**

Table 4: *National publication rates in selected languages for PCT milestone years*

Given this information, it is feasible to add the ***Table 4*** sub-totals (additional national publications produced in these four PCT publication languages) to the corresponding absolute numbers of PCT publications in the same languages at the same points in time (i.e. the data in ***Table 2*** expressed in absolute numbers rather than percentages). This presents us with a simple snapshot of the contribution of each language to the PCT Minimum Documentation in each year studied. This combination of data is shown in ***Table 5,*** and re-drawn in graphical form at ***Figure 5.***

Period	National patents from 1920 onwards (*)	International and regional systems
1970-1984	FR-A, -B, -U DE-A, -B JP-A, -B SU-A1, -A2, -A3, -A4 CH-B GB-A, -B US-A, -B + additional En, Fr, De basics without priority	WO-A 1978+ EP-A, -B 1978+ OA-B, -U 1982+ AP-B 1984+
1985-1998	as above + additional Es basics without priority	as above + EA-A, -B 1996+
1998-2007	as above + RU-B	
2007-2012	as above + KR-B	
2012-date	as above + CN-B	

Table 5: *Combined contribution of PCT and national publications to the PCT minimum documentation*

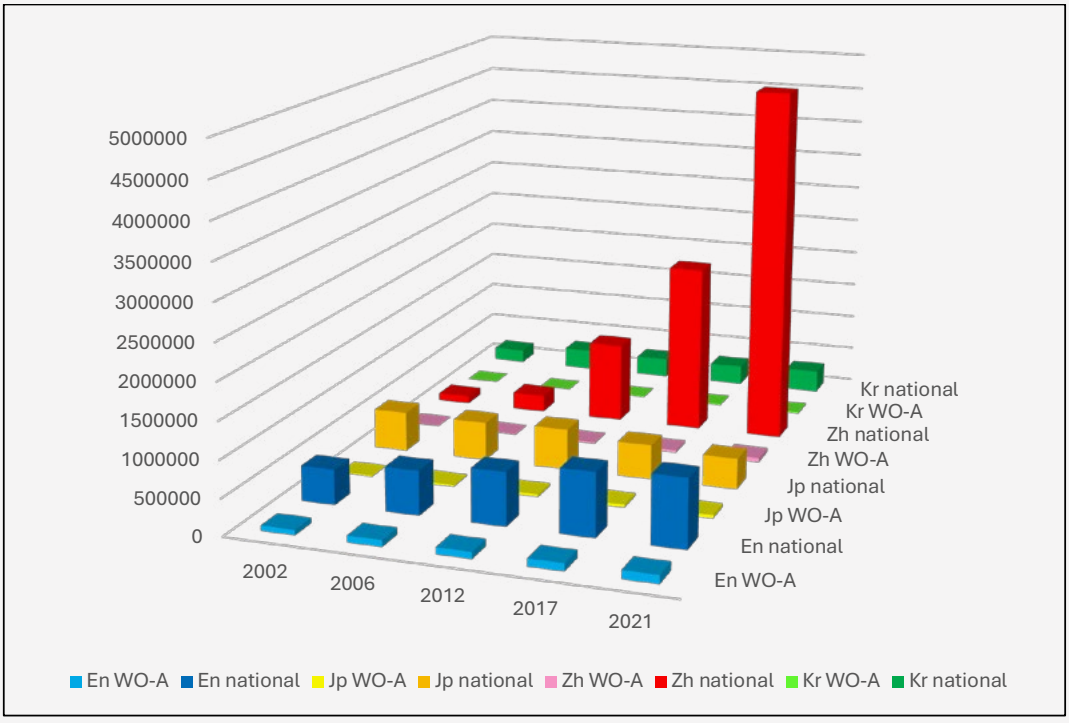

Figure 5: *Language distribution of national and PCT documentation, 2002-2021*

Conclusions?

It is immediately clear from this analysis that the rate of growth of national publications in Chinese has far outstripped the rate of growth of Chinese-language PCT publications, and the proportion of the entire PCT Minimum Documentation set which originates in English is being impacted. Analysis of the national authority files, using the appropriate Kind of Document codes and/or application number formats, may allow for the possibility of creating further data visualisations, analogous to Figure 4, which would permit a greater understanding of how the distribution of languages across the entire PCT Minimum Documentation has begun to drift decisively away from English. At some point, the ***lingua franca*** of the IP world may become Chinese, including our default text search language, and Chinese searchers will demand that global documentation is translated ***into*** Chinese, rather than ***from*** it.

About the author

Stephen Adams is the managing director of Magister Ltd., a UK-based consultancy specialising in patents information. Mr. Adams is a Qualified Patent Information Professional (number 20190044100092) and holds a B.Sc. in chemistry from the University of Bristol and an M.Sc. in Information Science from City University, London, as well as professional memberships of the Royal Society of Chemistry (RSC) and the UK's Chartered Institute of Library and Information Professionals (CILIP).

He is the author of three editions of "Information Sources in Patents", the latest published in 2020 by Walter de Gruyter KG, contributed several book chapters and written numerous articles in the field of patent information, including over 25 refereed papers for the Elsevier journal "World Patent Information".

His professional service includes the Editorial Advisory Board of "World Patent Information" between 2006-2020 and three terms on the Board of PIUG Inc., the International Society for Patent Information, as Director-at-Large (2002-2006) and Vice-Chair (2014-2016 and 2016-2018), as well as service on the management committee of the UK's Patent and Trade Mark Group over many years. He received the PIUG's Special Recognition Award in 2008 and the IPI Award in 2012 for outstanding contribution to patent information.

Tupolevlaan 81, 1119 PA Schiphol-Rijk
The Netherlands
Telephone +31 85 800 0024
info@lighthouseip.com

lighthouseip.com

The world's most complete IP data collection

LIGHTHOUSE IP