PATENTSCOPE の化学構造検索の検証（Ⅱ）[1]
～英語、日本語、中国語、韓国語の化合物名称での検索との比較

2022 年 1 月

アジア特許情報研究会：石川彰

【要旨】 農薬５種の化合物について PATENTSCOPE の化学構造検索と英語、日本語、中国語、韓国語の名称によるテキスト検索（以下、キーワード検索）の結果を比較しました。その結果、英語名称で検索ヒットした公報のうち 90%以上が化学構造検索でもヒットしました。一方、日本語、中国語、韓国語の名称での検索に関しては、化学構造検索でヒットできる割合が 10～40%と低くなりました。一部の化合物では、この割合が高いものもありました。本報告ではこの理由についても考察しました。

以上の結果より、PATENTSCOPE で日本、中国、韓国への出願公報を化合物から検索する場合には、英語名のテキスト検索だけでなく、日本語名、中国語名、韓国語名でのテキスト検索を併用したほうが良いと分かりました。

§1 はじめに

WIPO の提供する PATENTSCOPE に化学構造検索の機能が付いてから、約5年経過しました[2]。発表時には、無償のデータベースで化学構造検索ができるということが大いに注目されました。当時の対象国は WO、US で、インターフェースの言語、検索言語は英語でしたが、現在は言語も増えて、マーカッシュ検索や非特許文献の検索も可能になっています[3,4,5]。

筆者は、2019年に、本研究会ウェブサイトに簡単な検索検証の結果を投稿しました[1]。その検証では、WO 出願について化学構造検索と英語キーワード検索の結果を比較しました。ある農薬 Amidosulfuron について、英語キーワード検索でヒットする公報の 90%以上が化学構造検索でヒットしました。まずまずの結果と言えますが、理論上は 100%となってほしいところでした。

本報告では、検索能力の向上を確認するため前報[1]と同様な検証を行いました。検索対象化合物は農薬 5 種とし、キーワード検索の検索言語は、英語、日本語、中国語、韓国語としました。

§2 検証方法

§2-1 検証した化合物

検証に用いた化合物の化学構造式、及び、InChIKey（IUPAC 等で頻繁に使われる化合物の記述方法、InchI は International Chemical Identifier の略）、名称を Figure-1、Table-1 に示します。検索物質の日本語、中国語、韓国語の名称は、Google 翻訳の結果、及び公報中の使用状況を見て決めました。

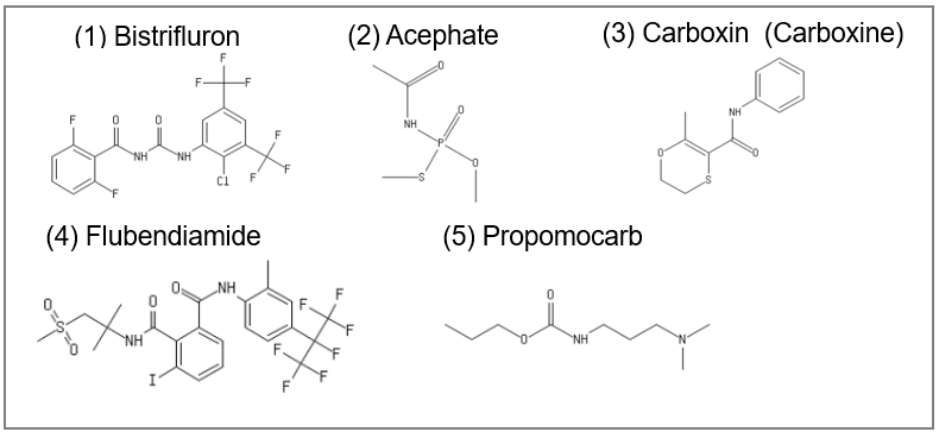

Figure-1　Chemical structures of compounds analyzed in this report（検証した化合物の化学構造）

Table-1 InChIKey and names of compounds analyzed in this report（検証した化合物の InChIKey・名称）

	Chemical compounds	Names of compounds			
	InChiKey	English	Japanese	Chinese	Korean
1	YNKFZRGTXAPYFD-UHFFFAOYSA-N	Bistrifluron	ビストリフルロン	双三氟虫脲	비스트리플루론
2	YASYVMFAVPKPKE-UHFFFAOYSA-N	Acephate	アセフェート アセフエート	乙酰甲胺磷	아세페이트
3	GYSSRZJIHXQEHQ-UHFFFAOYSA-N	Carboxin Carboxine	カルボキシン	萎锈灵	카르복신
4	ZGNITFSDLCMLGI-UHFFFAOYSA-N	Flubendiamide	フルベンジアミド	氟虫双酰胺	플루벤디아미드
5	WZZLDXDUQPOXNW-UHFFFAOYSA-N	Propamocarb	プロパモカルブ	霜霉威	프로파모카르브

※構造検索のコマンド　　：　CHEM:(InChIKey)
※キーワード検索のコマンド：　ALLTXT:(キーワード)

§2-2　PATENTSCOPE の検索条件

データベースは PATENTSCOPE を使用し[6]、US、JP、CN、KR の 4 つの出願国について比較しました。前報告と同様[1]、化学構造検索とキーワード検索とを比較しました。化学構造検索は、完全一致検索としました。

キーワード検索では、検索物質の名称を英語、日本語、中国語、韓国語としました。この物質名称でヒットする公報は、理屈上では化学構造検索の検索でもヒットすべきものと考えられます。上述 Figure-1 の各化学構造に関し、Figure-2 に示す集合A、B、C、その部分集合 1～7 の検索を行い、解析に使用しました。

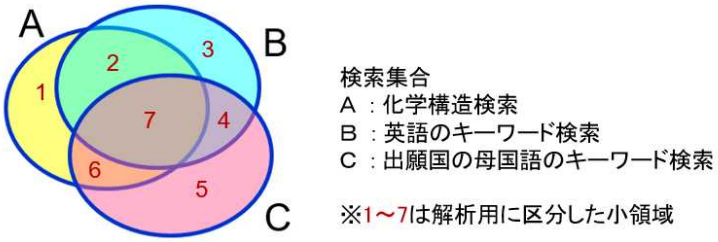

Figure-2 Venn Diagram used in analyses of this report（検討した検索集合の Venn 図）

検索モードは、化学化合物検索(Chemical compounds、以下、本文では「化学構造検索」と呼びます)のほかに、詳細検索(Advanced search)、構造化検索(Field combination)も使用しました。クエリの一例を下記に示します。

【農薬 Bistrifluron のクエリ例】

・キーワード検索
　英語:　ALLTXT:(bistrifluron)
　日本語:　ALLTXT:(ビストリフルロン)
　中国語:　ALLTXT:(双三氟虫脲)
　韓国語:　ALLTXT:(비스트리플루론)
・化学構造検索:　CHEM:(YNKFZRGTXAPYFD-UHFFFAOYSA-N)
・和集合の例　CHEM:(YNKFZRGTXAPYFD-UHFFFAOYSA-N) OR ALLTXT:(bistrifluron)
・その和集合を中国出願に絞る場合
　CTR:(CN) AND (CHEM:(YNKFZRGTXAPYFD-UHFFFAOYSA-N) OR ALLTXT:(bistrifluron))

本報告では、つぎの点に関する検討結果を述べます。

1）化学構造検索と英語名称に関するキーワード検索とを比較

2）化学構造検索と出願国の出願国語の名称に関するキーワード検索とを比較

§3 結果と考察

§3-1 英語の化合物名称でのキーワード検索との比較

Table-1 の農薬 6 種について、化学構造検索とキーワード検索のヒット件数を比較しました。 Figure-3 は、その結果を出願国の米国、日本、中国、韓国の別に（以下、それぞれ US、JP、CN、KR）図示したものです。Figure-3 の縦軸(A*B)／B は、化学構造検索でヒットした公報群(A*B)の英語キーワード検索でヒットした公報群 B に対する割合です。

Figure-3 では、緑色矢印、水色矢印でマークした公報群を除き、縦軸値は 90%以上と高い割合になりました。しかし、理屈上は 100%となるべきところですので、改良は必要と思われます。

ここでは、縦軸値が低めになった原因検討として、次の二点(ⅰ)、(ⅱ)を検討しました。

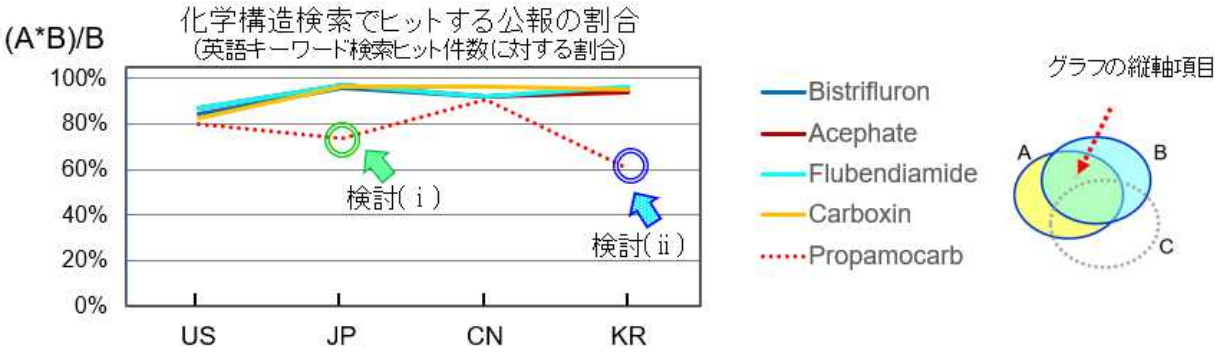

Figure-3 Ratio of publications hit by Chemical structure search against those hit by English keyword search. Applied country: US, JP, CN and KR.

（化学構造検索でヒットする公報数の割合（英語キーワード検索のヒット件数に対する割合））

検討(ⅰ):Figure-3 の緑色矢印で指示した公報群で縦軸値が低い理由（JP 公報）

検討のフローと結果を Figure-4 に示します。英語キーワードでヒットする B のうち、化学構造検索でヒットしない公報群 368 件（B not A）の中から任意の 10 件を選択し、化学構造検索でヒットしない理由を確認しました。

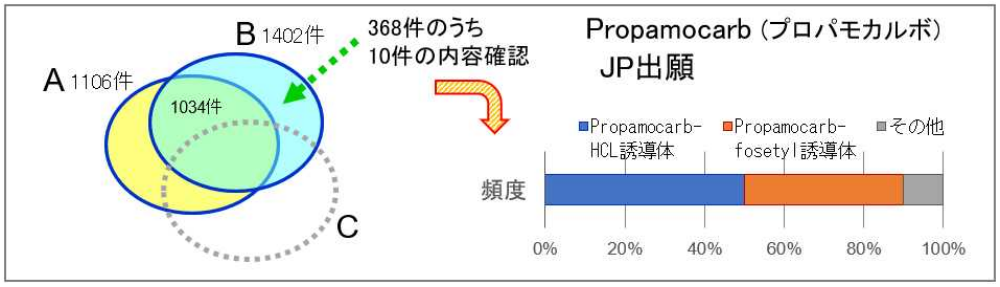

Figure-4 Procedures and results to analyze the reason why vertical axis values of Figure-3 are relatively low for search of “Propamocarb” of JP patents

（上述 Figure-3 の縦軸値が低い理由の検討フローと結果、検索対象:”Propamocarb”、出願国:JP）

Figure-4 の横棒グラフより、英語キーワード検索でヒットしながら、化学構造検索でヒットしない公報群の 90%が Propamocarb の誘導体と分かりました。誘導体は、次のようなものがありました。

誘導体の記載例:プロパモカルブ塩酸塩(propamocarb-hydrochloride)、

propamocarb-fosetylate(propamocarb-fosetylate)、プロパモカルブ-ホセチル(propamocarb-fosetyl)、プロパモカルブ-ホセチレート(propamocarb-fosetylate)

これらの誘導体については、化学構造検索では索引化が出来なかったと考えられます。

検討(ⅱ):Figure-3 の水色矢印で指示した公報群で縦軸値が低い理由(KR 公報)

Figure-5 に Figure-4 と同様な検討フローと結果を示します。

Figure-5 の横棒グラフから、化学構造検索でヒットできなかった公報群から任意に選択した 10 件中 9 件が Propamocarb 誘導体(HCL 塩)と分かりました。これらの誘導体が化学構造検索でヒットできなかった理由は、誘導体프로파모카르브염산염 (propamocarb hydrochloride)が、基本物質の프로파모카브 (propamocarb) と区別ができず、別の化合物として検索されたためと推定されます。

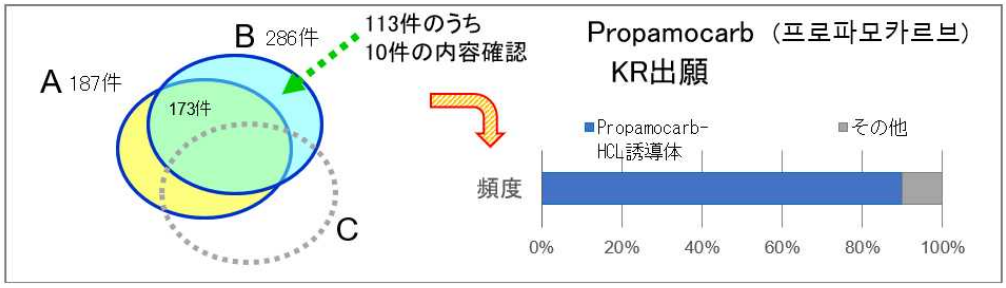

Figure-5 Procedures and results to analyze the reason why vertical axis values of Figure-3 are relatively low for search of “Propamocarb” of KR patents

(上述 Figure-3 の縦軸値が低い理由の検討フローと結果、検索対象:”Propamocarb”、出願国:KR)

以上のように、検討(ⅰ)、検討(ⅱ)より、該検索物質 Propamocarb については、誘導体を索引化できないと分かりました。今後、AIの利用等で改良されることを期待します。

§3-2 出願国語キーワード検索と化学構造検索との比較

PATENTSCOPE への化学構造検索導入当初は、検索キー抽出は、公報の英語部分から行われていましたが[1]、最近の発表では、カバーレッジも WO 及び IP-5(EP,US,JP,CN,KR)まで可能とあります(文献 4 の FAQ 資料)。公報の大部分は出願国の言語(以下、出願国語と呼ぶ)で出願されますので、出願国語の化学構造の名称検索でヒットする公報が化学構造検索でもヒットすることが望ましいと思われます。そこで、出願国語の名称検索と化学構造検索とを比較しました。

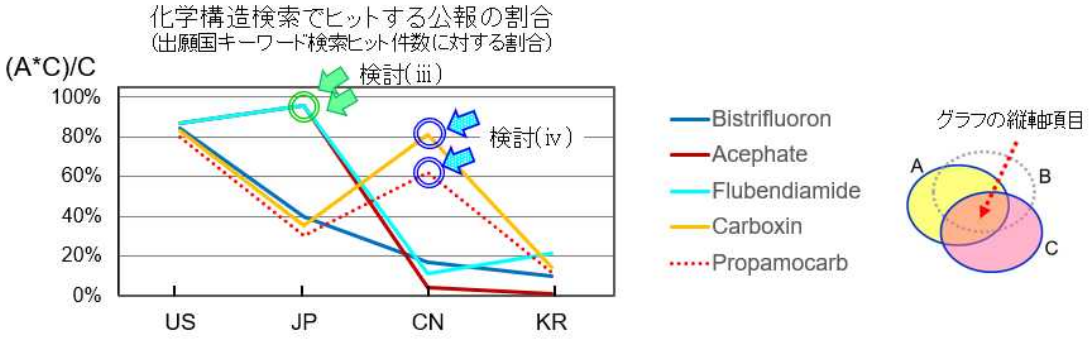

Figure-6 Ratio of publications hit by Chemical structure search against those hit by English keyword search, Applied country: US, JP, CN and KR.

(化学構造検索でヒットする公報数の割合(出願国語のキーワード検索のヒット件数に対する割合))

Figure-6に、出願国語キーワード(名称)でヒットした公報群Cのうち、化学構造検索でヒットする公報群(A*C)の割合(A*C)/C をグラフ化しました。ここで、US は、Figure-3 と同じデータです。緑色矢印、水色矢印

を除きますと、JP、CN、KR の(A*C)/C の縦軸値は、0%～40%であり、Figure-3 の US（英語）の場合と比べ低くなりました。一方、矢印の４点については、縦軸値は高くなっています。この違いがでる理由の検討として次の検討(ⅲ)～(ⅴ)を行いました。

検討(ⅲ)-1　Figure-6 の緑色矢印のうち赤線の Acephate(アセフェート)の縦軸値が高い理由

検討のプロセスと結果を Figure-7 に示します。Figure-7 によると、化学構造検索で、任意に選択した 10 件全件において、日本語"アセフェート”で化学構造検索用の索引化がされていました。

ここで、索引化されていると述べましたが、その例を Figure-8 に示します。"アセフェート"の語句にオーバーマウスすると化学構造がポップアップされます。これにより索引化の有無を確認できます。Figure-6 の緑色矢印の公報群では、任意に選んだ 10 件の全件に対して索引化を確認できました。

索引化されているため、化学構造検索でヒットしたと考えられます。このように、化合物によっては、出願国語でも索引化がされているものがあることを確認できました。

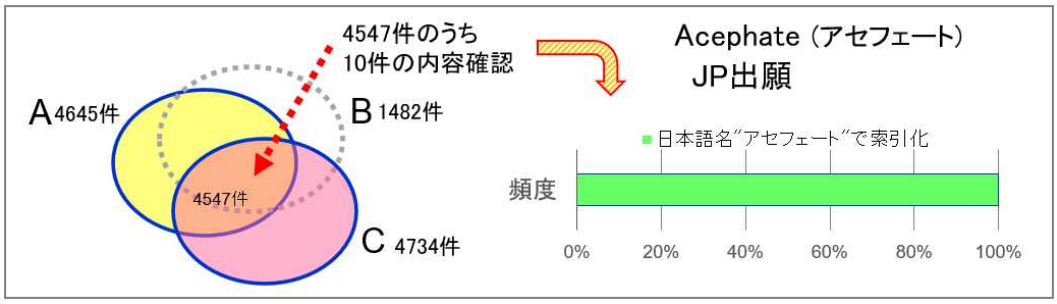

Figure-7　Procedures and results to analyze the reason why vertical axis values of Figure-3 are higher for search of “Acephate アセフェート” of JP patents
（上述 Figure-6 で縦軸値が高い理由の検討フロー、検索対象:”Acephate アセフェート”、出願国:JP）

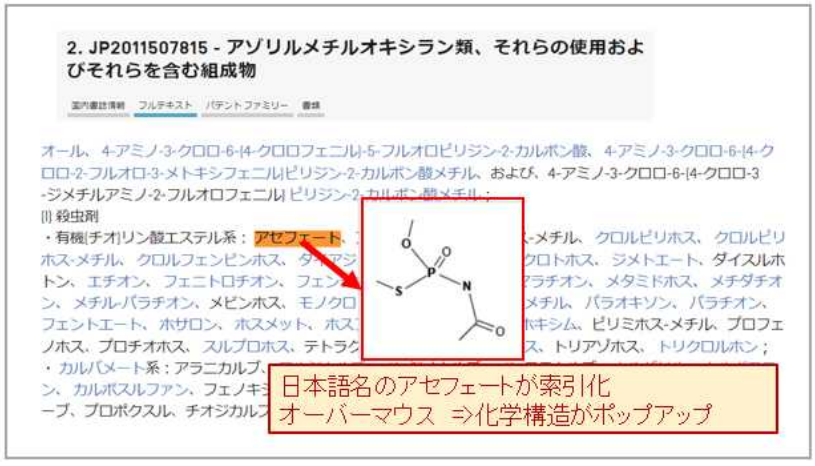

Figure-8　Example of indexing for Japanese word “アセフェート”
（索引化の例、日本語”アセフェート”の例）

検討(ⅲ)-2：Figure-6 の緑色矢印のうち水色線 Flubendiamide (フルベンジアミド)の縦軸値が高い理由

Figure-9 に検討のフローと結果を示します。Figure-9 の横棒グラフによると、任意に選んだ 10 件全件で"フルベンジアミド" が索引化されており、そのため、Figure-6 の縦軸値が大きめになったと考えられます。本例では、英語名称も並記されていました。

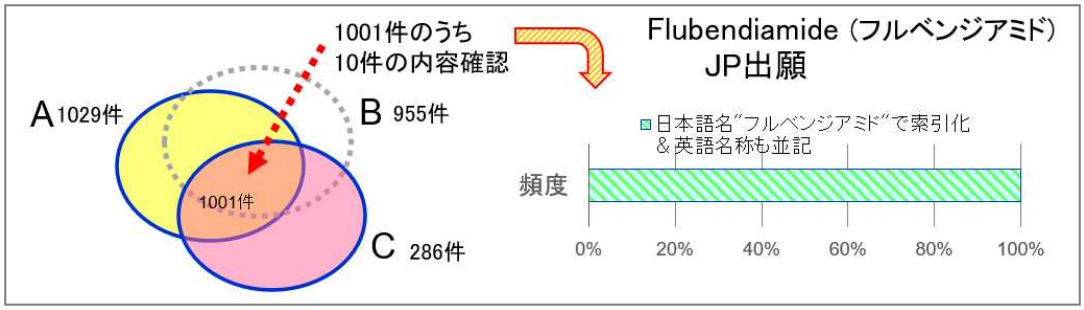

Figure-9 Procedures and results to analyze the reason why vertical axis values of Figure-3 are higher for search of “Flubendiamide フルベンジアミド” of JP patents

（上述 Figure-6 で縦軸値が高い理由の検討フロー：”Flubendiamide フルベンジアミド”、出願国：JP）

検討(ⅳ)-1：Figure-6 の水色矢印のうち黄色線 Carboxin（カルボキシン）の縦軸値が高い理由

Figure-10 に検討フローと結果を示しますが、中国語”萎锈灵”でヒットした公報群(C)のうち、化学構造検索(A)でヒットした公報群から選んだ 10 件について、ヒットした理由を調べました。Figure-10 の横棒グラフから、10 件全件で”萎锈灵”に化学構造検索用の索引化がされています。そのうち 40%に英語名の並記もありました。

そのため、Figure-6 の縦軸値が高くなったと考えられます。

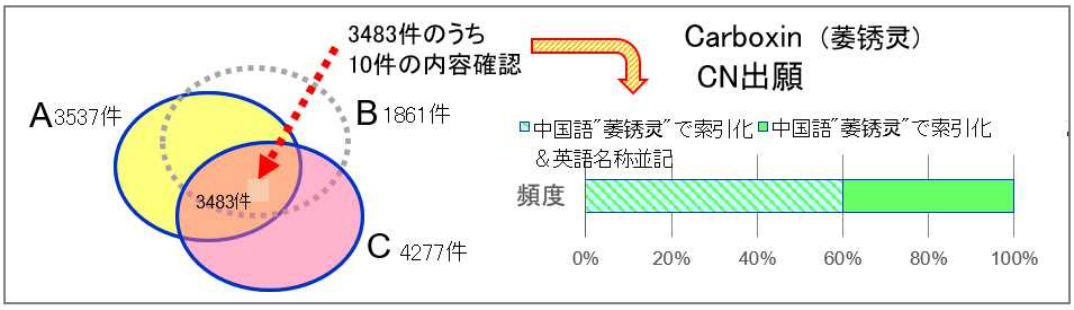

Figure-10 Procedures and results to analyze the reason why vertical axis values of Figure-3 are higher for search of “Carboxin 萎锈灵” of CN patents

（上述 Figure-6 で縦軸値が高い理由の検討フロー：” Carboxin 萎锈灵”、出願国：CN）

検討(ⅳ)-2:Figure-6 の水色矢印のうち赤色点線 Propanocarb(霜霉威)の縦軸値が高い理由

Figure-11 に検討フローと結果を示しますが、確認した 10 件のうち、70%で中国語"霜霉威"で化学構造検索用の索引化がされておりました。残り 30%では、英語名の記載がありました。 これらが原因で Figure-6 の縦軸値が高くなったと考えられます。

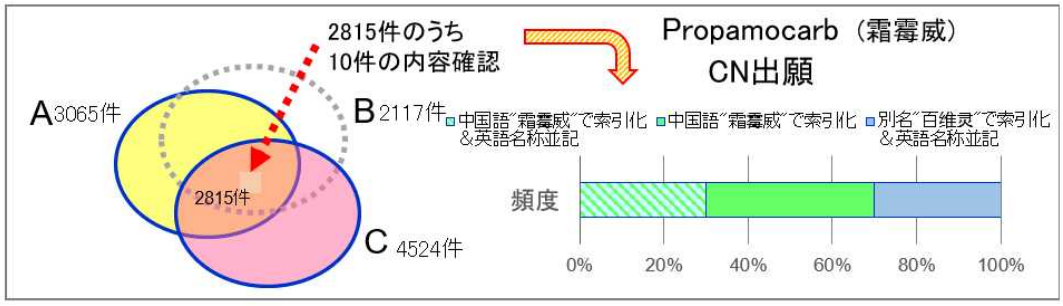

Figure-11 Procedures and results to analyze the reason why vertical axis values of Figure-3 are higher for search of "Propomocarb 霜霉威" of CN patents

(上述 Figure-6 で縦軸値が高い理由の検討フロー:" Propomocarb 霜霉威"、出願国:CN)

検討(ⅴ):

Figure-6 のうち、上述の検討(ⅲ)~(ⅳ)の化合物を除き、縦軸に(A*C)/C を横軸(B*C)/C に対して描いた図を Figure-12 に示します。ここで、

縦軸(A*C)/C:出願国語キーワード検索(名称)でヒットした公報のうち化学構造検索でヒットした公報の割合

横軸(B*C)/C:出願国語キーワード検索(名称)でヒットした公報のうち英語キーワード検索でヒットした公報(英語名称が並記されている)の割合

Figure-12 より、縦軸の(A*C)/C 値と横軸(B*C)/C 値とは相関関係があると考えられます。英語の名称記載の割合が高ければ、化学構造検索でヒットする割合が高くなると考えられます。並記された英語名称があるため、化学構造検索でヒットしたと考えられます。

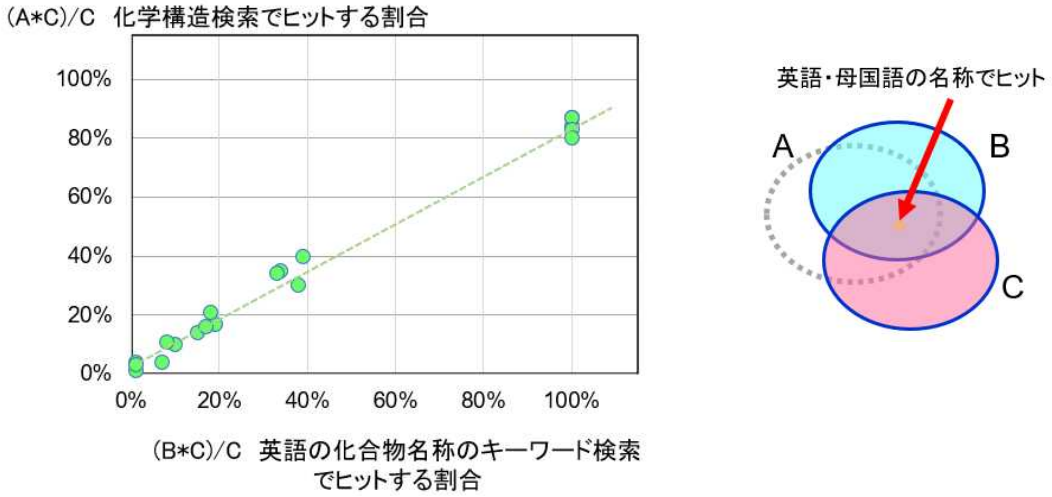

Figure-12 Relation between publications hit by keyword search of application country and those hit by keyword search in English

(化学構造検索でヒットする割合と英語キーワード(名称)検索のヒットする割合の関係)

§4 まとめ

PATENTSCOPE の化学構造検索で農薬 5 種について化学構造検索の結果を化合物名称のキーワード検索と比較しました。出願国は US、JP、CN、KR とし、キーワード検索の言語は、英語、日本語、中国語、韓国語の 4 か国語としました。

検討の結果、次の点がわかりました。

① 英語キーワード検索の結果の 90%以上は、化学構造検索でもヒットする。

② 出願国 JP、CN、KR について、言語を日本語、中国語、韓国語でヒットした公報の大多数が化学構造検索でヒットできない。

・ヒットするものは、英語（並記）で記載された名称があるためヒットした。

・ただし、出願国語においても化学構造検索用の索引化がなされているものがある。

・出願国語での索引化が進むことを期待する。

③ ヒット件数の比較をすると、次になります。一部のデータを除き、

ヒット件数　(英語キーワード検索) ≒ (化学構造検索) ＜ (出願国語キーワード検索)

従って、検索漏れを防ぐには、化学構造検索だけでなく、出願国の言語でのキーワード検索（名称での検索）も併用したほうが良さそうです。

§5 参考文献

1) 石川彰 「PATENTSCOPE の化合物検索の検証 ～キーワード検索との比較」
アジア特許情報研究会設立 10 周年記念誌=第 3 部：ワーキング資料 (2019)
※本報告を「PATENTSCOPE の化学構造検索の検証（Ⅰ）」とします。
http://patentsearch.punyu.jp/asia/2018ishikawa.pdf

2) 株式会社 IP エージェント「IP ニュース」「PATENTSCOPE に化学構造」(2016)
https://www.ip-agent.biz/topics/detail.php?nid=82

3) 特許庁 「PATENTSCOPE （特許文献の無料グローバル・データベース）の使い方 平成30年度」(2018)
https://www.jpo.go.jp/news/shinchaku/event/seminer/text/document/h30_jitsumusya_txt/11.pdf

4)WIPO PATENTSCOPE WEBINAR 資料 (2021)
https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_webinar_patentscope_2021_30/wipo_webinar_patentscope_2021_30_presentation.pdf

5) 酒井美里 「WIPO PATENTSCOPE マーカッシュ構造検索の利用方法（特許検索）」(2021)
https://note.com/sakaimisato/n/n1cf91c42a60e

6) PATENTSCOPE ログインサイト
https://ipportal.wipo.int/

以上