化学物質名の異表記同定手法に関する考察 *

田中るみ子（学籍番号　200821665）
研究指導教員：石塚英弘（藤井敦）

1. はじめに

化学物質は構造式，結合表，名称など様々な形式で表現することができ，その中で名称は体系名，慣用名，商品名，略名など多様な表記を持つ．どの表記を使うかは書き手次第であり，情報共有の阻害や情報検索における漏れが生じる．そこで，化学物質名の異表記を同定する手法が必要である．

多様な表記を持つ化学物質を構造に変換して比較することができれば，異表記問題は解決する．しかし現状ではあらゆる名称を構造に変換することは技術的に困難である．そこで，本研究は異表記の現象に着目し，なぜ異表記が発生するかという原因を化学的背景から分析し，異表記問題の本質を探ることを目的とする．

2. 本研究の概要

本研究は，化学物質名の異表記問題を検討するため，まず，実際の文書に記載された化学物質名を抽出し，その記載が従来からある化学物質データベースの記載とどのように違っているかを分析する．ここで，「表層的な特徴」だけでなく「化学的な特徴」から分析し，この結果をもとに同定手法の開発に向けた指針について考察する．

分析対象の文書として特許電子図書館 [1]から公開特許公報を検索して利用した．特許公報には，表記や翻訳に関する法的拘束がないため，研究者，技術者，弁理士など書き手の属性が多様であるという特徴から物質名の多様な表記がある．化学物質名の多様な記載が見られるため国際特許分類「C07D」を検索条件として指定した．「C07D」は複素環式化合物に関する物質名が多く記載されており，慣用名と体系名の組み合わせによって，名称が漸進的に増えるという特徴がある．

次に特許公報から抽出した化学物質名とそれと同一物質である化学物質が既存のデータベースではどのように記載されているかを調査した．既存のデータベースとして日本化学物質辞書 Web[2]（日化辞）を選択した．

特許公報と日化辞データベースを用いて同一物質でありながら記載が違う異表記対の集合（以下「異表記コーパス」）を作成し，異表記が生じた原因を目視で分析する．さらに，異表記コーパス中の事例を類型化する．類型化にあたり，物質名の表記的特徴に加えて物質そのものの化学的特徴を考慮する．

次に異表記同定手法に向けた考察を行う．具体的には異表記コーパス中の事例を構造に変換し，構造変換できないもしくは構造が一致しない異表記対はどのように同定すべきか検討する．

3. 異表記コーパスの作成

異表記コーパス作成にあたり，同一物質に対する異表記を特定することが非専門家には難しいという問題がある．そのため同一物質と共通に付与されている物質番号に着目した．本研究ではCAS登録番号を化学物質抽出の手がかりとして用いた．

日化辞では検索項目に「CAS 登録番号」があるため，CAS 登録番号を手がかりに抽出した物質と同一物質の別名称一覧を取得できる．一覧を見て，特許公報の名称と類似の名称を探し，CAS登録番号ひとつひとつに対して手作業で異表記対になる名称を探し，コーパスを作成した．

表 1 にコーパス作成用特許公報の検索条件を示す．公報発行日が 1993 年 1 月から 2009 年 9 月まで，国際特許分類が「C07D」複素環式化合物，「CAS 登録」が本文に含まれるという条件で検索を行い，公報 313 件を得た．表 2 に示すように，公報 313 件から CAS 登録番号の異なり 978 件を抽出し，これをもとに異表記対 201 件を作成した．

---

* “A Study of Identification for Chemical Substance Names” by Rumiko TANAKA

表 1 コーパス作成用特許公報の検索条件

公報発行日	1993 年 1 月～2009 年 9 月
国際特許分類	C07D 有機化学 複素環式化合物
公報全文	「CAS 登録」が本文に含まれる

表 2 異表記対作成に関するデータの件数

公報件数	313
CAS 登録番号の異なり	978
異表記対	201

4. 異表記の類型化

作成したコーパスの事例を一つ一つ見ながら，異表記の類型化を行った．まず表層的な特徴に着目し，表層的には説明がつかない異表記対は化学的特徴に起因すると考え，試行錯誤しながら，見直し，整理統廃合を行った．表層的特徴に類型化した中で，化学的背景を持つ場合は化学的特徴にも分類した．分類の結果を以下に示す．

・表層的特徴

命名方針，位置番号，立体表記法，記号の記載法，異表記，慣用名，説明語，記載順

・化学的特徴

異表記，字訳，翻訳，略語，誤り，不明

5. 異表記同定手法に向けた考察

化学物質を一意に特定する表現法として構造式がある．そのため，まず異表記コーパス中の事例ごとに名称から構造に変換し，比較した．次に構造変換できない，もしくは構造が一致しない異表記対はどのように同定すべきか検討した．

日本語名を「化合物名の和英翻訳 [3]」を用いて英語に翻訳し，Reaxys（リアクシス）[4]と OPSIN（オプシン）[5]を個別に用いて名称構造変換を行った．

名称構造変換の結果を表 3 に示す．異表記対 201 件のうち複数の分類に属する対は重複して計上したため異表記対の合計は 226 件となった．

「構造一致」は異表記対の両方が構造に変換でき，かつ両者の構造が一致した件数を表す．「構造不一致」は構造変換ができたものの両者の構造が一致しなかった件数を表す．「構造変換できない」は対のうち少なくとも片方が構造変換できなかった件数を表す．表 3 の合計欄に示すように Reaxys において「構造一致」は 24%，「構造不一致」は 11%，「構造変換できない」は 65%であった．OPSIN はそれぞれ 4%，0%，96%であった．

仮に構造変換と類型の間に強い相関があれば，類型ごとに構造変換や他の手法の提案ができる．しかし，表 3 の結果からそのような相関は得られなかったため，類型化別ではなく全体として考察する．

表 3 名称同定に構造変換を適用した例

分類		Reaxys			OPSIN		合計
		構造一致	構造不一致	構造変換できない	構造一致	構造変換できない	
表記的特徴	異表記	2	1	1		4	4
	字訳	21	1	11	1	32	33
	翻訳	2	1	4		7	7
	略語	1	3	3		7	7
	誤り	1		4		5	5
化学的特徴	命名方針	7	8	49	2	62	64
	位置番号	3	1	5	1	8	9
	立体表記			10		10	10
	記号の使い方	12	2	30	3	41	44
	異表記	1		4		5	5
	慣用名		2	8		10	10
	説明語		5	4		9	9
	記載順	4	1	14	2	17	19
合計		54	25	147	9	217	226
(%)		(24)	(11)	(65)	(4)	(96)	(100)

表層的特徴か化学的特徴かにかかわらず文字列の類似があることを考慮すると，文字の類似度比較に柔軟に対応できる文字単位および部分名称単位の N グラムが考えられる．化合物名称が化学的に意味のある部分名称に分けることができる特徴を利用すれば，部分名称の並べ替えが考えられる．しかし，本研究では計算機上で異表記を同定する手法を具現化することはできなかった．

6. おわりに

本研究の成果は，異表記問題を検討するため，実際の文書から異表記対を集めコーパスを作成し，異表記対の類型化を行った点にある．残された課題は，コーパス作成用の収集データに偏りがないか検討することと，異表記同定手法を確立することである．

注

1) http://www.ipdl.inpit.go.jp/homepg.ipdl
2) http://nikkajiweb.jst.go.jp
3) http://homepage1.nifty.com/nomenclator/chemjtra/chemjtra.htm
4) https://www.reaxys.com/
5) http://wwmm.ch.cam.ac.uk/wikis/wwmm/index.php/Oscar3