複単語表現と機械翻訳

首藤公昭　福岡大学名誉教授　工博
研究工房ことばの森　http://jefi.info

複単語表現MWEとは？

- 慣用句、コロケーション、決まり文句、およびこれらに類似の表現、ことわざ、格言、成句、複合語、固有表現（named entity)など、まとまった言語的機能を持つが要素語の共起に特異性(idiosyncracy)を持つ単語列

- 複単語表現 MWE： Multiword Expression
  --- I. Sag, et al., 2002 “Multiword Expressions: A pain in the neck for NLP”

  … → MWE-WN 2019、EUROPHRAS2019、MUMTTT2019、etc.…

- MWU、lexical bundles、formulaic language、construction grammar、phraseology、慣用連語

現状

- MWEは現在もMTの弱点の一つ
- MWE-annotate された大型コーパスが必要、現状では一部の固有表現、複合語、VPC、LVC、SVCを除き、テキスト上で多様なMWEを同定(identify)することができない
- 多数の低頻度MWEがコーパスから捉えにくい(ジップの法則)

→　広範な表現に対応し、構文情報、変化形情報をもった大型MWEレキシコンが必要

--- C. Ramish 2017 “Putting the Horses Before the Cart: Identifying Multiword Expressions Before Translation” MUMTTT2017

--- A. Savary, et al. 2019 “ Without lexicons, multiword expression identification will never fly: A position statement” EUROPHRAS2019

-- PARSEME shared taskの総括

JMWELとは?

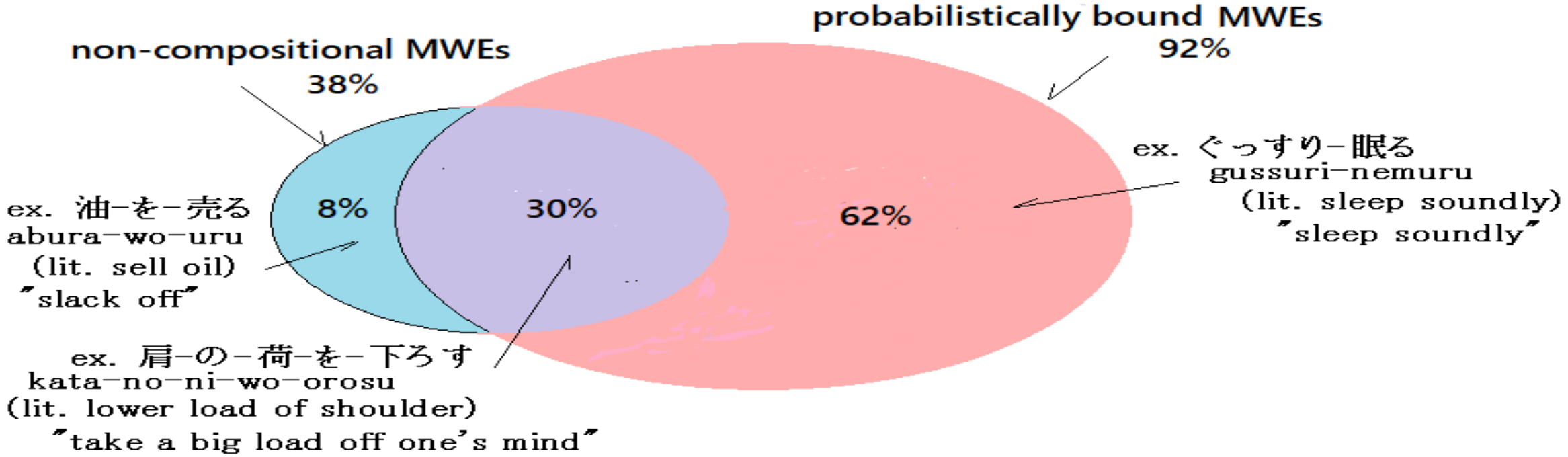
- JMWEL: Japanese MWE Lexicon は現在約159,000 見出しを持つ総括的日本語複単語表現レキシコン、固有表現を除く書き言葉現代日本語対象、 形態・構文的な機能および構造情報を詳細に記載
- K. Church, 2011 “How many multiword expressions do people know?”, invited talk, workshop on MWE, ACL meeting への回答

- 1968年、ルールベースのPBMT開発を目標に編集に着手、以後継続的に増補改訂して現在に至る

- 電気科学技術奨励賞2011、言語資源賞2018

JMWELの表現採録基準

1. 非構成性(Non-compositionality)
2. 要素語間の確率的な縛り(強共起性: collocationality)



Rough Sketch of MWE distribution

文法的サブレキシコン　文法範疇を網羅、互いに素

1. **名詞性** MWEs　28,400　>　institutionalized phrases, non-compositional compound-nouns
2. **動詞性** MWEs　76,800

   2.1 class1 (36,400)　基本的トライグラム　N-p-V　　>　　LVCs, SVCs

   2.2 class2 (36,200)

   2.3 class3 (4,200)　　=　　non-compositional compound-verbs
3. **形容詞性** MWEs　5,800
4. **形容動詞性** MWEs　2,800
5. **副詞性**　MWEs　17,600
6. **連体詞性** MWEs　17,100
7. **接続詞性** MWEs　1,900　=　complex discourse markers,　complex sentence adverbs
8. **機能語性** MWEs　7,800

   8.1 post-predicative (5,100) =　complex auxiliary verbs,　complex ending-particles

   8.2 postpositional (2,700) =　complex case-particles,　complex connective-particles

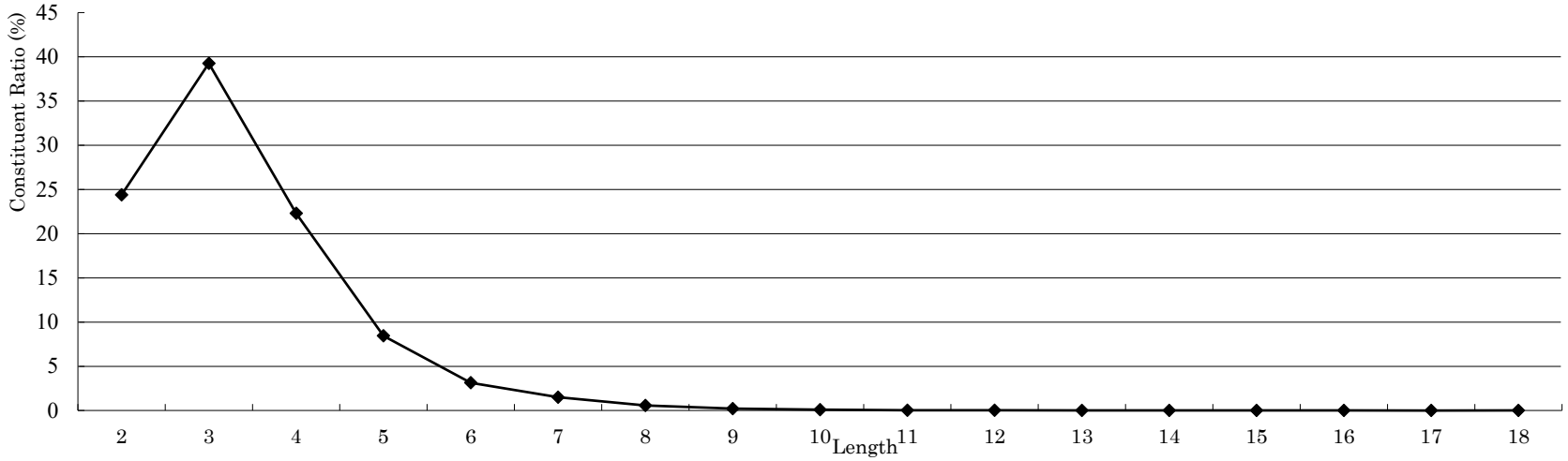
トピック的サブレキシコン

1. **標準的慣用句** 4,900
2. **オノマトペ共起表現** 43,600
3. **四字熟語** 3,200
4. **不完全句** 470
5. **ことわざ/決まり文句/格言/成句** 4,000

---

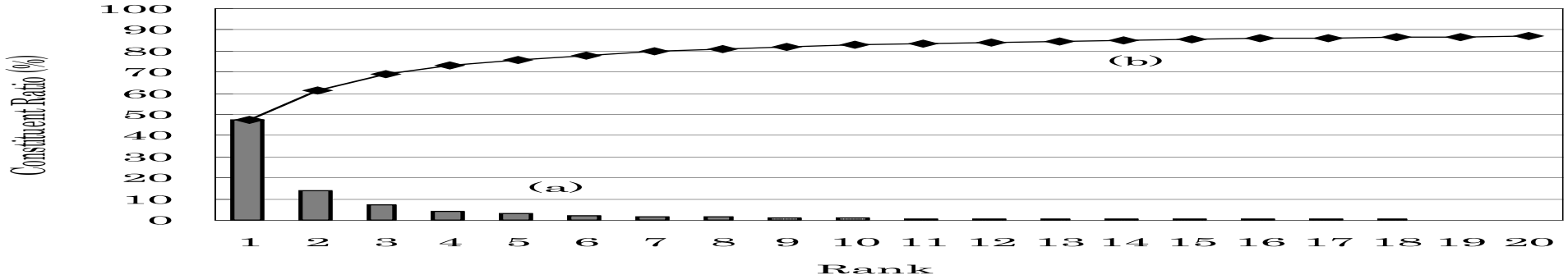
6. **標準的慣用句の英訳例文レキシコン** 4,000 ----- under construction
7. **呼びかけ/応答/挨拶/独言 /間投表現** 1,000

長さの分布

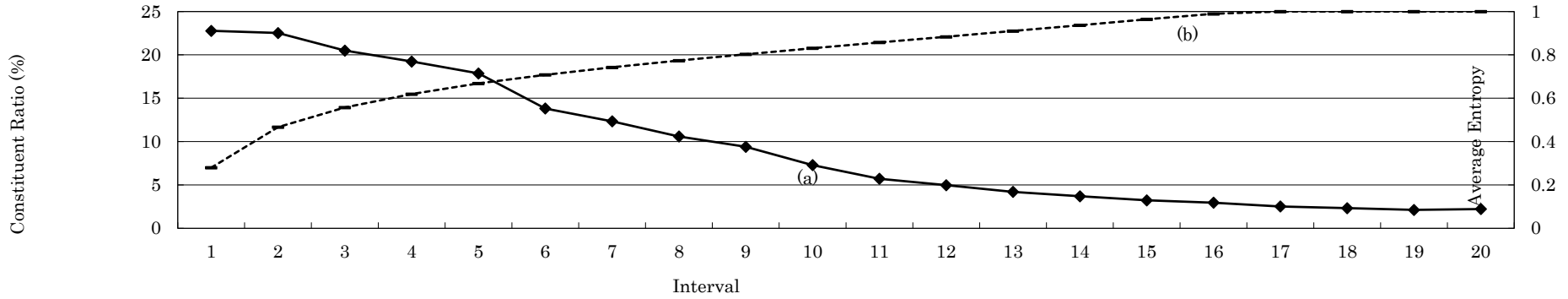

Constituent Ratio (%)
45
40
35
30
25
20
15
10
5
0
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Length

収録の妥当性 --- 200億文webコーパス上のn-gram 頻度データ LDC2009T08による検証 --- 動詞性class1(NpV)の場合

遷移確率　p(V|Np)



正規化エントロピー　H'(V| Np)

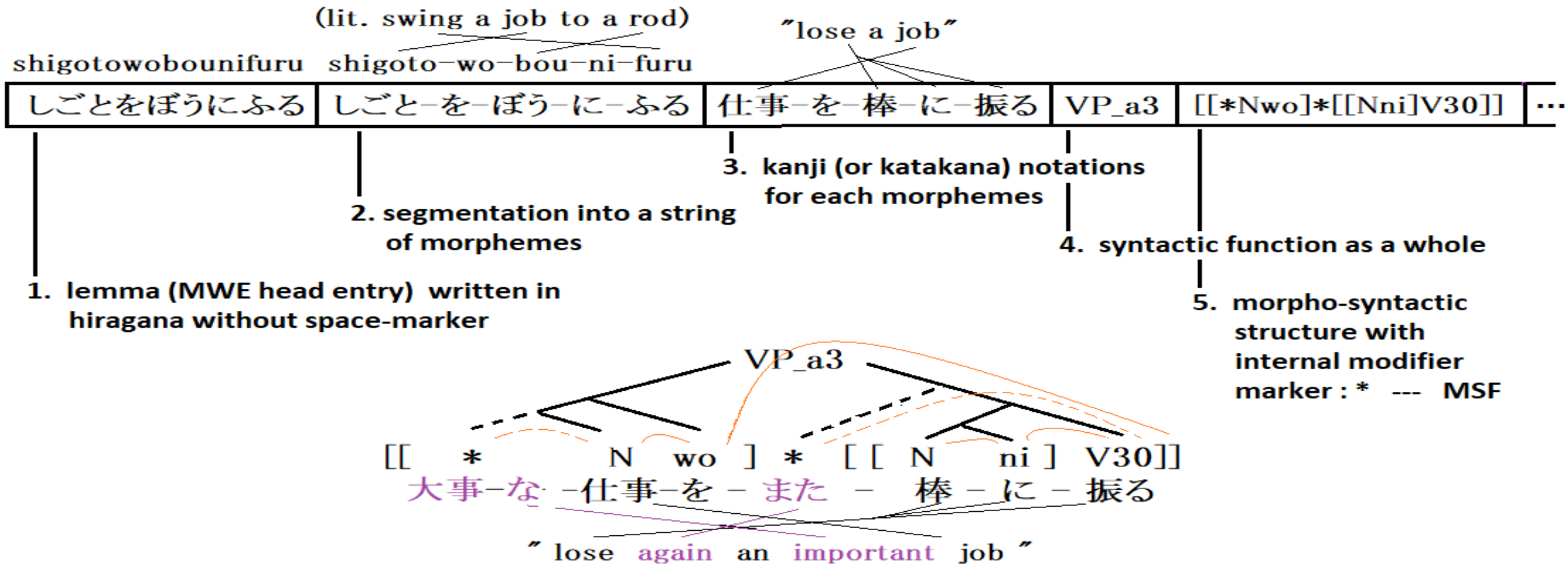


標準的慣用句の英訳例文レキシコン

segmented lemma	notational variation	function	morpho-syntactic structure	Japanese examples of usage	English translation
ご_たく-を-ならべる	御_託-を-並べる	VP_V	[Nwo]V30	御託を並べるな	1. Don't [lecture/preach to/harp on at] me.
					2. None of your silly talk.
こころ-が-ときめく	心-が-(ときめく/トキメく)	VP_V	[*Nga]V30	彼に初めて会った時、彼女は心がときめいた	Her heart was pounding when she first met him.
こころ-に-うかぶ	心-に-浮かぶ	VP_V	[*Nni]V30	1. そのとき、彼のとけるような笑顔がふと心に浮かんだ	At that time his beautiful smile came to mind.
				2.突然ある考えが心に浮かんだ	Suddenly a certain idea [occurred to me/hit me/flashed into my mind/came to mind/came to me].
こころ-を-うた-れる	心-を-打た-れる	VP_Vreru	[[Nwo]V15]reru	彼女は彼の無償の愛に心を打たれた	She was very [touched/moved/impressed] by his selfless love for her.
こころ-を-うばわ-れる	心-を-奪わ-れる	VP_Vreru	[[*Nwo]V15]reru	1. 彼はその女性に心を奪われた	He was fascinated by the woman.
				2. 彼はすっかり音楽に心を奪われている	He is completely enraptured by music.
こし-が-すわる	(コシ/腰)-が-(据(わ)/座)る	VP_V	[Nga]V30	彼は腰がすわらない	He can't [hold a job long/commit himself to anything].
ごじっ-ぽ-ひゃっ-ぽ	五十-歩-百-歩	NP_N	《([NN])([NN])》	1.（諺）五十歩百歩	A miss is as good as a mile.
				2. どっちが悪いかなんて言えないよ両方とも五十歩百歩なんだから	I can't say which is to be blamed since it's six of one and half a dozen of the other.
こし-を-すえる	(コシ/腰)-を-据える	VP_V	[Nwo]V30	1. 結婚したら東京に腰を据える	We are going to settle in Tokyo after we get married.
				2. 彼女は腰をすえて仕事をしている	She is [settled in/devoted to] her job.
ことば_じり-を-とらえる	言葉_尻-を-(捉/捕)える	VP_V	[*Nwo]V30	彼は人の言葉尻を捕える悪い癖がある	He has the bad habit of picking up on the petty points of others.

全サブレキシコン共通の基本情報、他の情報はサブレキシコン固有・多種

e.g. 仕事を棒に振る　　----　　大事な仕事をまた棒に振る



内部修飾可能性マーカー“*”の意義

1. ギャップ付き(不連続)MWEが扱える

　　＋　MWEの構文的柔軟性の度合い;

2.　通常句(free word combination)に近いMWEが扱える.

e.g.

[[Nni][[Nwo]V30]]　−−− **syntactically rigid MWE　　e.g. 手-に-汗-を-握る**

[[*Nni][[Nwo]V30]]　[[Nni]*[[Nwo]V30]]　[[Nni][[*Nwo]V30]]　[[Nni][[Nwo]*V30][

[[*Nni]*[[Nwo]V30]]　[[*Nni][[*Nwo]V30]]　[[*Nni][[Nwo]*V30]]　　…

[[*Nni]*[[*Nwo]V30]]　[[*Nni]*[[Nwo]*V30]]　[[*Nni][[*Nwo]*V30]]　　…

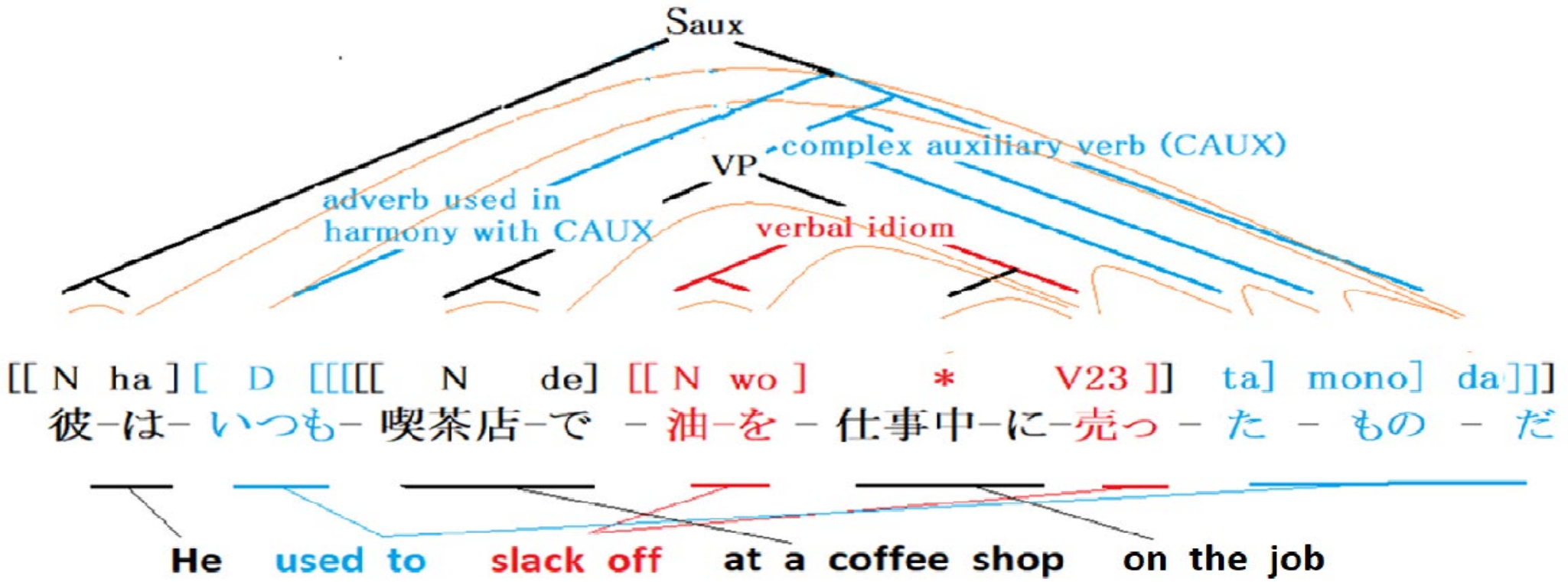
[[*Nni]*[[*Nwo]*V30]]　−−−　**syntactically flexible expression (free word combination)　e.g. 足-に-マメ-を-作る**

3. “*”のない場所に修飾句があれば、そのMWE候補は通常句　　**--->**　***semantic disambiguation***

テキスト上でMWEを同定(identify)できることが重要

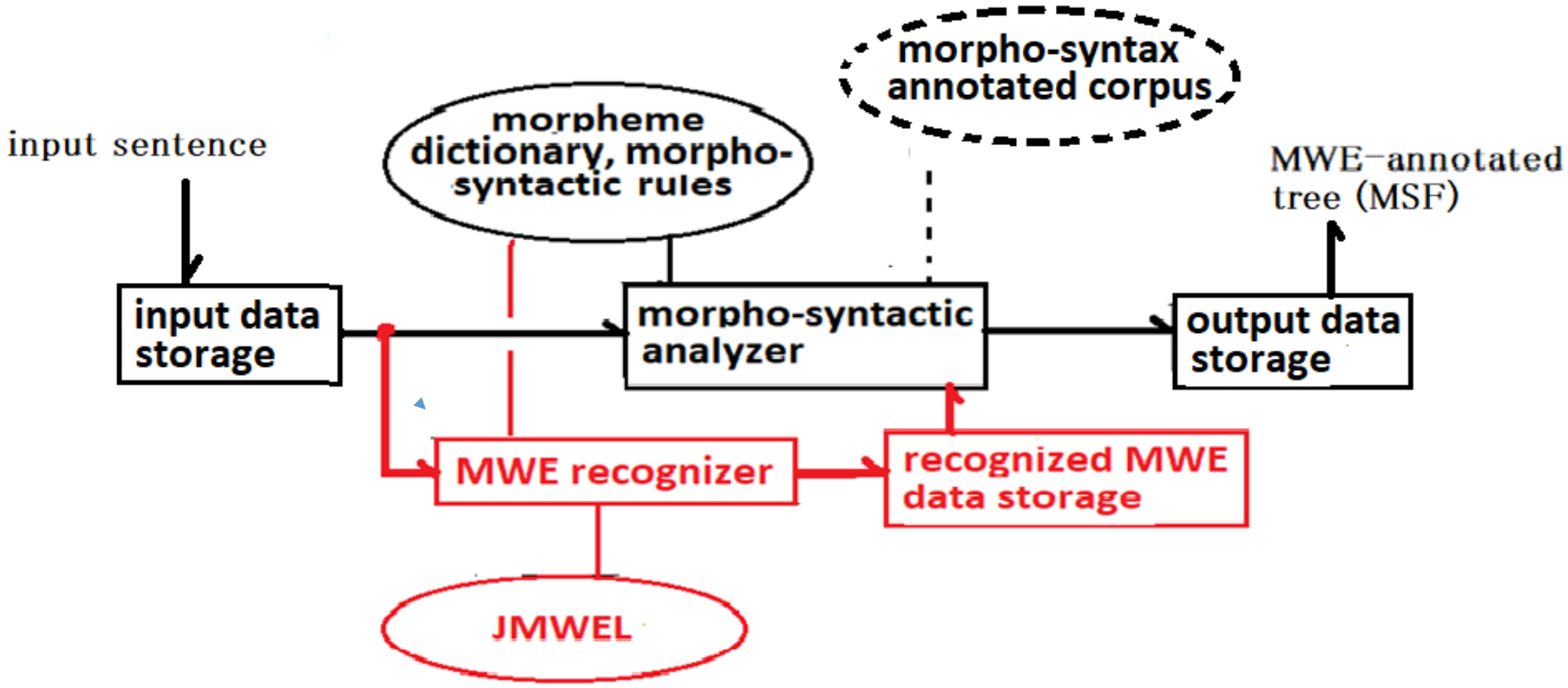
e.g. **油を売る** [[Nwo]*V30] **Verbal Idiom** --- Eng. **slack off**

たものだ [[[$ ta]mono]da] **Complex AUX** ＋ left context: **いつも** --- Eng. **used to**



MWE-based parser

pat. 5379318



課題

- 日本語コーパス上でMWE-アノテーション
  → MWE-annotated コーパスの開発

- 構成的MWEと非構成的MWEの仕分け
  → 非構成的MWE(トークン)の内容語マーキング
  → マークした語の文脈ベクトルとマーク無し同一語文脈ベクトルと
  の差異＝MWEの意味的非構成性の度合い

- MWEのシンタクス情報等でphrase tableのスコア付けを補強したtree-to-string あるいは tree-to-treeタイプの syntactical SMTの実現

ご清聴を感謝します

ご質問や利用方法等については viggo_ksf@jcom.home.ne.jp へ

または http://jefi.info からどうぞ