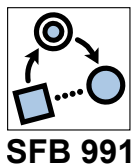
Mehrworteinheiten

Lexikale Kondierung: Lichte et al. (to appear)

Timm Lichte

HHU Düsseldorf

7. Sitzung, 28.11.2017



Überblick

Was heißt “to appear”?

- draft → manuscript (ms.) → submitted (sub.) → accepted (acc.), forthcoming → to appear (ta.) → in press → published (YEAR), in print, out of print

Aufbau:

- Introduction
- On the notion of regularity
- The most basic encoding format
- General virtues of lexical encoding formats
- Challenges posed by MWEs
- Fixed MWE encoding formats
- Fully-flexible encoding formats
- Summary

Introduction

what is it that makes an encoding format suitable for encoding multi-word expressions (MWEs) as part of an electronic resource?

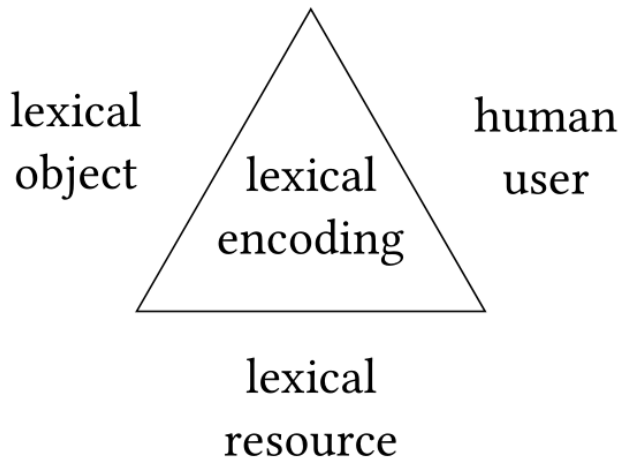


Figure 1: Interface aspects of lexical encoding

- regularity versus irregularity
- general aspects and typical examples of encoding formats

On the notion of regularity

Given: E = set of objects, p = property

irregular / idiosyncratic

p is shared by **exactly one** member in E.

regular property

p is shared by **at least two** members in E.

trivially regular property

p is shared by **all** members in E.

non-trivially regular property

p is shared by **at least two but not all** members in E.

On the notion of regularity (cont.)

Given: E = set of objects, e = object in E, P = property set

regular object

e has a property set P with only regular properites as to E.

Degree of regularity:

- likelihood of a property set
- diversity of property sets

This notion of (ir)regularity implies that it is impossible to determine once and for all whether the properties of certain objects are regular or irregular, simply because the set of conceivable properties and objects is unbounded. In other words, the whole business of telling apart regularity from irregularity hinges on the selection of properties along with a specific set of objects.

On the notion of regularity (cont.)

Meanings of “irregular” wrt. MWEs:

1. the set of **objects** is sufficiently **restricted**
   (e.g., by contrasting the MWE with non-MWEs only);
2. the set of **properties** is sufficiently **extended**
   (e.g., by taking into account very specific properties of the MWE);
3. the **property set** of the MWE is relatively **unlikely** and
   “irregular” is assigned a likelihood related meaning.

$\Rightarrow$ high risk of overlooking or neglecting some regularities

The most basic encoding format: Property name sets

(1) a. kick-the-bucket :=
{NP_0 V NP_1, NP_1.Det.the, NP_1.N.bucket, V.kick, meaning=die}

b. spill-beans :=
{NP_0 V NP_1, NP_1.N.beans, V.spill, passive, meaning=divulge}

Interpretation function from property names to objects of target formalism (z.B. Bäume, Strings, Merkmalsstrukturen, XML)

notational adequacy how close the encoding format is related to the target formalism

The most basic encoding format: Property name sets

(1) a. kick-the-bucket :=
$\{NP_0\ V\ NP_1, NP_1.Det.the, NP_1.N.bucket, V.kick, meaning{=}die\}$

b. spill-beans :=
$\{NP_0\ V\ NP_1, NP_1.N.beans, V.spill, passive, meaning{=}divulge\}$

Advantages:

(i) it is very **flexible** in terms of adding and removing property names and adapting the interpretation function to some target formalism;

(ii) it makes **empirically largely neutral** descriptions available;

(iii) it is **conceptually lean and inviting** for formal novices because the main data structures are just ordinary sets.

But:

- nobody would seriously make use of property name sets when encoding a large electronic lexicon

The most basic encoding format: Table encoding

Table 1: Table encoding of the property name sets in (1)

ID	NP_0 V NP_1	NP_1.det	NP_1.N	V	passive	meaning
kick-the-bucket	+	the	bucket	kick	-	die
spill-beans	+		bean	spill	+	divulge

$\Rightarrow$ Lexicon-Grammar framework (Gross 1994)

General virtues of lexical encoding format

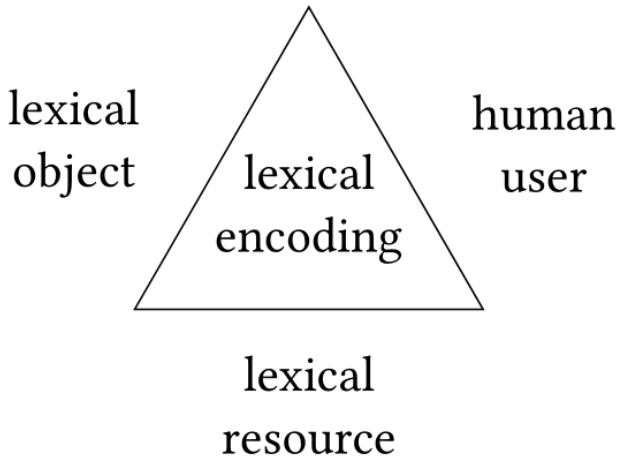


Figure 1: Interface aspects of lexical encoding

Virtues as to the lexical object

- completeness: every property is uniquely mapped onto a property name (injective)
- conciseness: the property name set is minimal (surjective)

(Table encoding in Table 1 ist not concise!)

General virtues of lexical encoding format

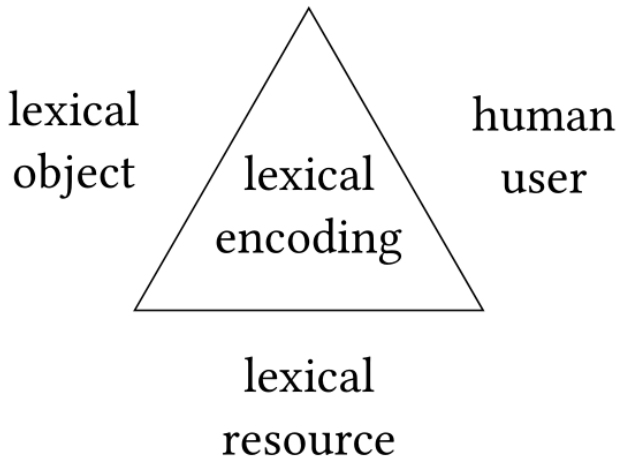


Figure 1: Interface aspects of lexical encoding

Virtues as to a human user

- transparency: cognitively reversible mapping
- flexibility: allows the user to freely choose property names and to include new properties on the fly
- power to generalize: the parts of encodings should be reusable at any level of representation and detail
- implementation friendliness:

General virtues of lexical encoding format

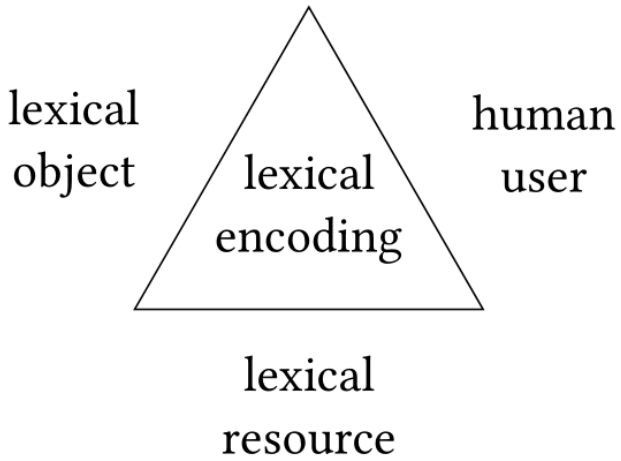


Figure 1: Interface aspects of lexical encoding

Virtues as to the lexical resource

- electronic versatility: the relative ease with which a corresponding lexical encoding can be converted into a lexical resource

Challenges posed by MWEs

Among others:

- restrictive agreement: (EN) *to cross one's fingers*
- defective subcategorization, i.e. imposing a subcategorization frame which the MWE headword does not admit outside MWEs, e.g. (PL) *dobrze mu z oczy patrzy* (lit. well him looks from eyes) 'he looks like a good person' prohibits a subject: */uczciwość dobrze mu z oczy patrzy/ (lit. honesty well him looks from eyes), while *patrzy* 'looks' as a standalone verb always requires one

(i) regularity of properties of MWEs is scale-wise

(ii) properties of different degrees of regularity co-occur in each MWE

(iii) truly idiosyncratic properties are rare,

(iv) shared properties can be unforeseen

Fixed MWE encoding formats

fixed = tailored to specific regularities and languages

Beispiele:

- DuELME (**Gregoire2010 Gregoire2010**) fürs Niederländische
- Walenty (Przepiórkowski et al. 2014) fürs Polnische

Fixed MWE encoding formats: DuELME

Dutch Electronic Lexicon of Multiword Expressions: 5000 Einträge

```
1  % Pattern description
2  PATTERN_NAME ec1
3  POS d n v
4  PATTERN  [.VP [.obj1:NP [.det:D (1) ] [.hd:N1 (2) ]] [.hd:V (3) ]]
5  DESCRIPTION Expressions headed by a verb, taking a direct object
       consisting of a fixed determiner and a modifiable noun.
6
7  % MWE description
8  EXPRESSION zijn kansen waarnemen
9  CL zijn kans[pl] waar_nemen[part]
10 PATTERN_NAME ec1
11 EXAMPLE hij heeft zijn kansen waargenomen
```

Figure 2: DuELME pattern description ec1 (from Grégoire 2007b) and MWE description of (NL) *zijn kansen waarnemen* (lit. *one's chances perceive*) 'to seize the opportunity' (from Grégoire 2010)

Fixed MWE encoding formats: Walenty

Walenty: Valenzlexikon mit 100 000 Einträgen

```
1  patrzeć: np(dat)+advp(misc)+lex(prepnp(z,gen),pl,'oko',natr)
```

Figure 3: Description of *dobrze [KOMUŚ] z oczu patrzy* (lit. *well someone*.DAT *from eyes looks*) 'someone looks like a good person' in Walenty

Flexible MWE encoding formats

flexible = properties, property names and inference rules (or macros) can be freely chosen

Beispiele:

- PATR-II (Shieber (1986))
- XMG (Crabbé et al. 2013; Petitjean, Duchier & Parmentier 2016)

Flexible MWE encoding formats: PATR-II

- true classic, very influential
- descriptions of CFG rules with feature structures (aka. directed acyclic graphs)
- templates, lexical rules, default inheritance

```
18  Define SubjectPossObjectAgreement as
19      [subject: [agr: $1]
20      object: [poss: [agr: $1]]]
```

Flexible MWE encoding formats: PATR-II (cont.)

```
22  Define ZijnKansenWaarnemen as
23      Transitive
24      SubjectPossObjectAgreement
25      [lex: waarnemen
26       object: [lex: kans
27                agr: [num: pl]
28                modifiable: -]
29      sem: [paraphrase: seize_the_opportunity]]
```

Flexible MWE encoding formats: PATR-II

- more flexible than DuELME: defining properties, “pattern” factorization
- clear denotational semantics
- more theory dependent?

Still some inflexiblities:

- Templates only apply to “root node” of FS.
- FS are untyped.
- Word order constraints are difficult to encode.

Flexible MWE encoding formats: XMG

- generates a wide range of linguistic ressources
- dedicated compilers using adapted description languages
- object-oriented programming (encapsulation)

```
class subject_poss_object_agreement
declare ?Subj ?Obj ?NUM ?PERS ?GEND
export ?Subj ?Obj
{ <syn> {
    ?Subj[num=?NUM,pers=?PERS,gend=?GEND];
    ?Obj [] {
      [cat=d,num=pl,possnum=?NUM,pers=?PERS,gend=?GEND] "zijn"}}}
```

Flexible MWE encoding formats: XMG (cont.)

```
18 class zijn_kansen_waarnemen
19 import transitive[] subject_poss_object_agreement[]
20 declare ?I
21 { <syn> {
22     ?Subj[i=?I];
23     ?Obj [] {
24       [cat=n,modifiable=-,num=pl] "kans"};
25     ?V[] "waar_nehmen" };
26   <frame> {
27      [using-event,
28       actor:?I,
29       theme:chance]}}
```

Flexible MWE encoding formats: XMG (cont.)

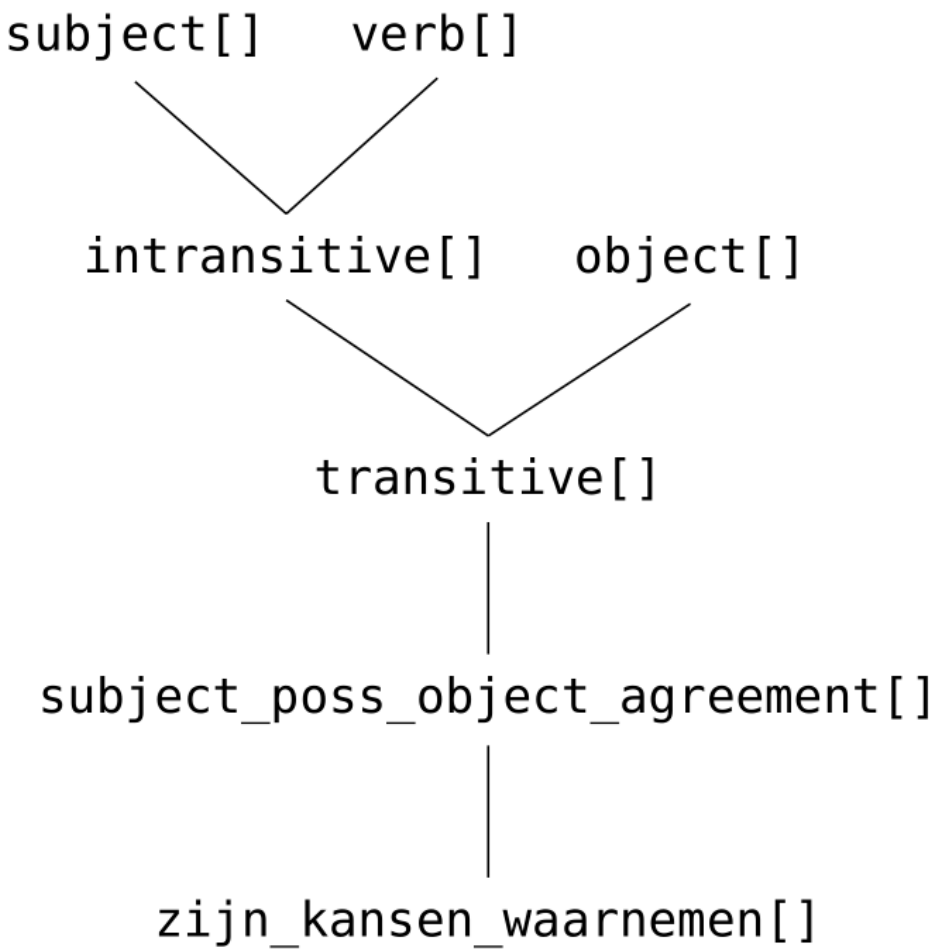


Figure 7: Inheritance hierarchy of XMG classes according to the code in Figure 6

Summary

	transparency	flexibility	power to generalize	implementation friendliness	electronic versatility
DuELME	4	4	3	2	1
Walenty	3	3	3	1	1
PATR-II	1	2	2	4	4
XMG	1	1	1	3	3

Maybe the most outstanding feature of many MWEs is their **semantic non-compositionality**, and addressing it in a lexical encoding framework remains one of the most challenging perspectives.

References

[1] Crabbé, Benoit, Denys Duchier, Claire Gardent, Joseph Le Roux & Yannick Parmentier. 2013. XMG: eXtensible MetaGrammar. *Computational Linguistics* 39(3). 1–66. http://hal.archives-ouvertes.fr/hal-00768224/en/.

[2] Gross, Maurice. 1994. Constructing lexicon-grammars. In Beryl T. Sue Atkins & Anotonio Zampolli (eds.), *Computational approaches to the lexicon*, 213–263. Oxford: Oxford University Press.

[3] Petitjean, Simon, Denys Duchier & Yannick Parmentier. 2016. XMG2: Describing description languages. In Maxime Amblard, Philippe de Groote, Sylvain Pogodalla & Christian Retoré (eds.), *Proceedings of Logical Aspects of Computational Linguistics (LACL) 2016, nancy, december 2016* (Lecture Notes in Computer Science 10054), 255–272. Berlin: Springer.

[4] Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski & Marek Świdziński. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international Conference on Language Resources and Evaluation, LREC 2014*, 2785–2792. Reykjavík, Iceland.

[5] Shieber, Stuart M. 1986. *An introduction to unification-based approaches to grammar.* (CSLI Lecture Notes Series 4). Stanford, CA: Center for the Study of Language & Information.