Anna Braasch
Center for Language Technology (CST), Denmark

There's no Accounting for Taste – Except in Dictionaries ...*

Abstract

This paper discusses essential aspects of a new methodology for a corpus based lexicographic approach as developed within the project on Descriptive Lexical Specifications (DELIS, LRE 61.034). The main topics treated here are (1) description and analysis of electronic corpus data in terms of semantic frames and (2) interpretation of the analysis results towards formal lexical representations to be used in computational lexicography. The frame semantics approach provides the theoretical basis for lexical specification of corpus data. The descriptive model (exemplified by perception verbs) interrelates semantic, syntactic and morphological levels.

0. Introduction

DELIS is a broadly formulated project covering 7 work packages with different working tasks. Research and development are carried out in co–operation between a number of institutes and language groups including Danish. The main objective of the project is innovative in itself: to integrate corpus exploration, lexical modeling and tool development. Key tasks are

- lexical description on the basis of observable facts in corpus material
- contrastive lexicology across languages within commonly selected semantic domains and on the basis of a common theoretical hypothesis
- use of a common descriptive model for lexical specification
- tool development for corpus search, formal modeling and dictionary encoding.

For each task the key notion is reusability. For a more detailed description, see Heid (1994) elsewhere in these proceedings.

This contribution mainly deals with items related to the three first mentioned tasks. For Danish there does not yet exist any dictionary or comprehensive research report containing systematic semantic descriptions for verb classes based on detailed analysis of corpus data. Electronic corpora are widely used in the ongoing work on the dictionary of contemporary Danish *(Den Danske Ordbog)* similar to corpus based dictionary work for English (e.g. in COBUILD). Also systematic investigations of lexical

semantics and verb classification are carried out for English (Levin 1993). However, there is still a need for elaboration and testing of a comprehensive strategy to link observed semantic, syntactic and morphological features into a description of a lexeme and the corpus evidence (sentence) it occurs in. The methodology is based on recent theoretical hypotheses on cognitive-frame semantics developed by Fillmore (a.o. in Fillmore 1992).

0.1 Project aspects to present

The first part of this paper concerns selected aspects of methodological and descriptive issues addressed by the current project. The main outcome so far is the elaboration of the Corpus Evidence Encoding Schema (Krüger and Heid 1993), a device for specification of well-formed linguistic objects and annotation of corpus sentences. The cross-lingual character of the project entails different interpretations of the descriptional guidelines.

The second part deals with the language specific application of the Corpus Evidence Encoding Schema (CEES) for Danish within the perception semantic domain. This topic is illustrated by the analytical description of corpus sentences containing the verb *smage* (taste) or *lugte* (smell).

The conclusion – which also forms a kind of 'accounting for taste' – points to observations from the perception corpus analysis relevant to bottom-up dictionary construction.

1. Methodology and approach

The theoretical framework for systematic lexical description is based on the following key notions:

- the lexical analysis has as a basis the cognitive-frames approach which is supported by recording the relevant syntactic and morphosyntactic properties based on an HPSG-like approach
- dictionary entries are to be based on corpus evidence.

1.1 Linguistic approach

The cognitive-frames approach has been elaborated and instantiated for the semantic domain of sensation and perception by Charles Fillmore. The basic assumption is that language expressions make explicit or implicit mention of the conceptual elements involved in realisation of environmental phenomena as sense impressions. In language manifestations these conceptual elements are expressed by semantic frame elements (FEs) that also are called semantic roles. Although the prototype language is English, the underlying theory deals with language independent linguistic phenomena and it therefore can be successfully applied to other languages.

A draft paper (Fillmore 1992) offers a preliminary overview of the above mentioned semantic domain including a division into the modalities taste, smell, sight, hearing and touch. The further subdivision sketched out is based on the physical sense types, e.g. contact vs. distant senses (taste vs. smell), or chemical vs. non-chemical senses (taste vs. hearing). A number of illustrative examples showed that a given cooccurrence of verb + frame elements in the sentence constitutes the particular sense of the verb.

This draft has been reworked and extended within the framework of DELIS (cf. Fillmore et al. 1993) with guidelines for the description of the relevant context elements: semantic roles (e.g. experiencer, percept, judgment) and their grammatical and lexical realisations (cf. Zaenen 1993). In parallel, analyses of corpus examples has been carried out for Danish, Dutch, English, French and Italian to test the validity of the approach and the guidelines given so far (cf. e.g. Atkins 1993 and Braasch 1993).

The outcome of the working steps outlined above is a general descriptive format – the so called Corpus Evidence Encoding Schema (CEES) (Krüger & Heid 1993).

1.2 The CEES as descriptional device

"The main innovation, here, which is introduced by CEES, is that we regard the definition of CEES not only as an inventory of possible labels, but as a structured domain of definitions with rules determining well-formed descriptions of linguistic objects... " Heid states in the Main Report on DELIS (Ostler 1994:20).

The CEES serves the following purposes:

- to exploit corpus evidence of the keyword for lexical description
- to provide a standardized, computationally reusable descriptional format
- to correlate semantic, syntactic and morphosyntactic properties of the keyword and corpus sentence in a systematic and unambiguous way based on the underlying theory
- to provide a basis for polyfunctional lexical specifications of the keyword allowing generalisation.

The CEES records the observable features as attribute-value pairs. The main information types for a sentence containing the selected keyword (e.g. *smage, lugte*) are the following:

- sentence-level types, i.e. the corpus sentence itself, sentence properties and contextual information
- word-level types, i.e. semantic domain of the keyword and its frame element (FE-)group.

The FE–group reflects the cognitive and structural pattern of the keyword and it thus makes up the core of the lexical description. The frame element group is described by a set of features. For each context element included by the FE–group the following quadruple is recorded:

- Semantic role (or FE)
- Grammatical Function (GF) e.g. SUBJ, OBJ, IOBJ, COMPL
- Phrase Type (PHR–T) e.g. np, pp, advp, i.e. as usual in terms of noun, prepositional or adverbial phrases and clause/sentence types
- Lexical expression (EXPR), i.e. the particular piece of context.

This is a top–level generalisation of the descriptive guidelines. It needs appropriate language–specific adjustments and specialization according to both the part–of–speech and the semantic domain to be described. The process of refinement leads to operational CEES subtypes such as a CEES for Danish perception verbs.

1.3 Selecting an appropriate subcorpus

The largest available machine readable corpus for Danish is established by the Danish Dictionary (DDO), it contains about 40 mill. tokens. From this we first extracted a great number of corpus sentences for selected perception verbs (e.g. approx. 29.000 corpus examples of *se* (see)) and then the material to be analyzed was reduced to a manageable size. The selection process also included the sorting–out of domain external homographs, e.g. the case of the preteritum form *så* (saw) of the verb *se* , the homographs of which translate into English as 'so' (conjunction and adverb) or 'sow' (e.g. wheat). Because of the large amount of material, the number of corpus sentences for each keyword has been restricted to a maximum of 1,000 examples. The randomized sampling method (which selects every Xth concordance line) has been used although we also considered other statistically founded (e.g. frequency–based) methods.

1.4 Corpus evidence and word sense disambiguation

Recent work on sense disambiguation – also for dictionary encoding purposes – involves corpus analysis as a supporting aid. In DELIS, lexical analysis and description including word sense disambiguation starts (and ends) with corpus work: the properties of a keyword are captured as they occur/appear in the corpus. On the other hand, in a ready–made dictionary entry, corpus sentences are used as documentation e.g. of the keyword senses. In the following, we primarily refer to the work done on lexical analysis within the perception domain.

1.5 Analyzing corpus evidence

The subcorpus containing the selected perception verbs *smage* (taste) and *lugte* (smell) showed a wide variety and complexity of linguistic phenomena. The first action taken was to draw up a preliminary set of classification criteria based on observed syntactic patterns and the cognitive frame elements they realize. The overall meaning structure of a keyword is reflected by different sets and realisations of set of the frame elements appearing in the example sentences. (For convenience the term 'frame configuration' will be used to name a frame element set and its given realisation.) A dialectical shifting between syntactic and semantic aspects has turned out to be a fruitful investigation method.

As the next step in the process a first sorting of the corpus sentences into primary classes is carried out based on the established lexical semantic criteria. Further steps specialised and refined the descriptional criteria. In parallel, the lexical realisations of the relevant frame elements, such as of percept, experiencer and judgment were listed and divided into types based on their semantic content.

2. The application of CEES

In the analysis process the subcorpus has been roughly examined for keyword senses and structural patterns. The keyword senses were defined based on a monolingual Danish dictionary (*Nudansk Ordbog* 1990) and a bilingual Danish–English dictionary (Vinterberg & Bodelsen 1991). A subset of sentences was selected covering a wide range of syntactic patterns and representing all core senses of the verb that were found in the above–mentioned dictionaries. Sentences containing unusual verb senses, ad–hoc, idiomatic or metaphorical uses, etc. were copied into a separate file for later investigations. In parallel, we adapted the general CEES to the needs of Danish (including conceptual frames for English proposed by Fillmore) and elaborated the operational CEES's. These operational CEES subtypes were applied to corpus sentences within the perception domain. We tested the overall descriptional power of the methodology and approach and the usefulness of the coding guidelines.

The outcome of the corpus evidence encoding showed that the methodology and linguistic approach in general are also well–suited for Danish, although a few adjustments were needed, particularly within the language specific realisation of frame elements and in the morphosyntactic area.

2.1 An accounting for 'taste' and 'smell' in Danish

On the basis of the filled–in CEES (approx. 100 selected corpus sentences for each verb) we tabulated the frame element configurations occurring in the corpus material.

In Fillmore's terminology the verbs *smage* (taste) and *lugte* (smell) denote chemical experiences but the first one concerns a contact sensation, the second a distant one. A first investigation showed that the frame element configurations of the two verbs can differ according to the physical circumstances.

Levin classifies verbs of English on the basis of their syntactic behavior, i.e. their possible combinations of arguments and adjuncts in various syntactic expressions (Levin 1993:2), which are also known as alternations. The observed patterns of semantically determined syntactic properties (i.e. the possible alternations) lead to a verb classification where the class of 'see verbs' are made up by a number of perception verbs, which beside 'taste' and 'smell' also include e.g. 'detect' and 'discern'. On the other hand, 'taste' and 'smell' make up a subclass taking only a limited range of sentential complements compared to other 'see verbs'. In Danish, we observed a similar tendency.

2.2 Frame element configurations

In this paper we only point to a few essential observations from the semantic analysis. The examples below illustrate a number of fundamental findings wrt. frame element configurations.

The list of semantic roles in the sensation/perception domain at the top level is as follows: EXPERIENCER – PERCEPT – JUDGMENT. Their most frequent basic grammatical functions are SUBJECT – OBJECT – ADVERBIAL, respectively. However, the presence or absence of the EXPERIENCER (as active or passive role) affects the distribution of the grammatical functions within the sentence, which means that the logical object of the verb, the percept, functions as subject of the sentence (see also item (3)).

The notation used here follows a pattern in accordance with the quadruple defined by the CEES (cf. item 1.2), namely

FRAME ELEMENT NAME/type (GRAMMATICAL FUNCTION; phrase type) + FRAME ELEMENT NAME/...etc.

(The last member of the quadruple, the lexical realisation is omitted.)

(1) The EXPERIENCER/active (SUBJ;np) + PERCEPT (OBJ;np) configuration appears in three basic cases:

(a) the keyword is supported by a modal verb and expresses the faculty of perception;
(b) the keyword is used within extended or
(c) metaphorical senses:

(a) Jeg kunne lugte fiskene i kurven.
('I smelled the fish in the basket.')

(b) Anna har aldrig smagt slik.
('Anna has never tasted sweets.')
and: Han ønskede at smage friheden igen.
(Lit.: 'He wanted to taste freedom again.')

(c) ...et job jeg tidligere har lugtet lidt til.
(Lit.: '... a job I have smelled cursorily before' which means 'smattered to'.)

(2) The EXPERIENCER/active (SUBJ;np) + PERCEPT (OBJ;pp) configuration expresses attending. However, the valency bound prepositions are discerning for the investigated verbs:

(a) Han lugtede til pigens hår.
('He smelled [to] the hair of the girl'.)

(b) Hun smagte på suppen og ...
('She tasted [on] the soup and...')

(3) The EXPERIENCER/passive (i.e.not mentioned explicitly) + PERCEPT (SUBJ;np) + JUDGMENT (COMPL; advp or pp) configuration is typical when a sensory experience is evaluated (a) or interpreted (b):

(a) Suppen smagte fortrinligt.
('The soup tasted delicious.')

(b) Gamle mennesker lugter på en særlig måde.
(Lit.: 'Old people smell in a particular way.')

Closer investigations of the phrase type and lexical expression of FE's lead us to relevant language specific observations. Below a few examples:

(4) It turns out that the PERCEPT element appears with discerning subtypes with *smage* and *lugte*, respectively.

(a) 'Han lugtede fra munden.'
(Lit: 'He smelled from the mouth').

The frame element PERCEPT receives here the subtype label Source/ locus. The reason why no parallel structure pattern exists for *smage* can be seen in the conceptual difference between distance and contact senses.

(5) Investigating lexical realisations of the JUDGMENT frame element, we recorded a number of elliptic sentences, which means that the sensory quality adjective which evaluate the phenomenon (as adverbial complement) is absent. The corpus examples showed a consistent difference between the semantic content of the omitted JUDGMENT frame elements of *smage* {positive} and *lugte* {negative}, respectively.

(a) Du lugter!
(Lit.: 'You smell' {bad, awful}!)

(b) Ih, hvor det smager!
(Lit.: 'Oh, it tastes' {good, delicious}!)

In such cases the expressive function of interjections are recorded too.

(6) Many pieces of corpus evidence for the selected keywords are phrasal verbs e.g. *smage til* (to season).

(a) Hun smagte suppen til med lidt salt
('She seasoned the soup with some salt.')

(b) Han smagte løs af vinen
('He drank large quantities of the wine.')

Types of multi–word units (like in (b)) containing the keyword are at the present stage of the project not treated, but the occurrences are stored in a separate file for later analysis.

(7) In a number of sentences the keyword appears with a less usual structure pattern; for instance it includes the reflexive pronoun *sig* too. The following is an example of verbs of seeking, a subtype of perception verbs. The PERCEPT element receives in this case the subtype label Target.

(a) Dyrene lugter sig hurtigt frem til maden.
(Lit.: 'The animals smell themselves quickly through to the food' which means that they quickly find the food because of the smell of it.)

In such cases we consider if the expressed meaning belongs to the core senses within the perception domain. Often we recognize a metonymic extension or a non–literal use of the keyword. For instance, *smage* (taste) form a part of fixed or semi–fixed metonymic expressions with various senses, like 'hesitate', 'try', 'feel', etc. In addition, domain–external use of perception verbs seems to be rather frequent. These examples also are recorded but not yet described in detail.

2.3 The use of CEES as database formula

The corpus sentences are entered into the PARADOX database, where the CEES is used as database record format. This allows for consistent encoding, quick update and clear survey of the entered data. Furthermore, a database system supports very effectively systematic sorting of data types and extraction of cross–tabulated information types. In this way, we easily obtain lists of co–occurring frame elements, their syntactic patterns and lexical realisations. In addition, lexical realisations with common semantic features can be extracted from the encoded database records too. Also the language specific morphosyntactic features are catered for.

The use of a device like the PARADOX DBMS supports further work towards generalisation and formal modeling very well.

3. Results and perspectives

In the present phase of the project we carried out a systematic assessment of the validity of a new methodology on corpus–based lexical descriptions and the proposed frame–semantics approach.

A simplified and generalised overview of frame configurations for *smage* and *lugte* is shown in Table 1.

VERBS OF	EXPERIENCER	PERCEPT	JUDGMENT
ATTENDING	ACTIVE	AIM	ZERO
PERCEIVING	PASSIVE	SOURCE/STIMULUS	ZERO
OBSERVATION	ACTIVE	INTERPRETATION	ZERO
SENSORY QUALITY EVOKING	(PASSIVE)	SOURCE/STIMULUS	PERCEPT QUALITY

SENSORY EXPERIENCE EVALUATION	(PASSIVE)	SOURCE/STIMULUS	EVALUATING
SENSORY EXPERIENCE INTERPRETATION	(PASSIVE)	SOURCE/STIMULUS	INFERENCE
SUSTAINED ATTENTION	ACTIVE	AIM	ZERO
SEEKING	ACTIVE	TARGET	ZERO

Table 1. Frame element configurations for *smage* and *lugte*

Each frame element, e.g. PERCEPT, can be subclassified and labelled according to more precise distinctions, such as for the stimulus–type percept the labels 'discriminatum' (i.e. a particular part or feature of the percept) and 'locus' (i.e. the location of the percept) can be used. The same frame element may be expressed with various syntactic categories. Furthermore, each category has a more or less delimited number of lexical realisations.

The testing phase provided a necessary and sufficient basis for the following working issues:

- further steps of classification and generalisation of the observed phenomena as prerequisite for hierarchically structured descriptional models
- large–scale encoding extended to cover an additonal semantic domain (speech acts)
- hierarchical structuring of lexical descriptions wiithin selected domains
- adaptation of the monolingual description models to contrastive lexicography.

Acknowledgement

* The author whishes to thank the project co–ordinator of DELIS, Ulrich Heid, Stuttgart, for his always very useful advice.

References

(papers internal to the DELIS project are marked with *)
Atkins, B.T.S. 1993. *Corpus Evidence Encoding Schemata for English.* Oxford.*
Braasch, A. 1993. *Corpus Evidence Encoding Schemata – Subtypes for Danish.* Copenhagen.*
DELIS – Main Report. 1994. See: Ostler, N.
Fillmore, Ch. 1992. *A Cognitive–Frames Approach to the Vocabulary of Sensation and Perception in English.* Berkeley. *

Fillmore, Ch., B.T.S. Atkins and N. Ostler 1993. *Recommended Coding Categories.* London/Stuttgart. *
Heid, U. 1994. *Relating lexicon and corpus: Computational support for corpus–based lexicon building in DELIS. In: EURALEX '94.* Amsterdam.
Krüger, K. and U. Heid 1993. *On the DELIS Corpus Evidence Encoding Schema (CEES).* *
Levin, B. 1993. *English Verb Classes and Alternations.* Chicago: The University of Chicago Press.
Ostler, N. (ed.) 1994. *DELIS – Main Report. 1994.* Deliverable D–II of DELIS (LRE 61.034). London.*
Zaenen, A. 1993. *First report on syntactic tagging in DELIS.* Pisa/Stuttgart. *

Dictionaries

Nudansk Ordbog. 1990. Politikens Forlag A/S. Copenhagen .
Vinterberg, H. & C.A. Bodelsen. 1991. *Dansk–engelsk ordbog.* Gyldendal. Copenhagen.