Systematic Polysemy of Nouns and its Lexicographic Treatment in Estonian

Margit Langemets
Institute of the Estonian Language, Tallinn, Estonia

The focus of the study of the polysemy of the Estonian noun (Langemets 2010) was on identifying the systematic patterns of noun polysemy with further perspective to elaborate the principles to encode and represent systematic polysemy of nouns in the database of the one-volume dictionary of Estonian (to appear in 2015) and in the **ƎELex** *(= EELex) dictionary management system of the Institute of the Estonian Language.*

The analysis was based on the lexical perspective, i.e. on the lexicographic representation of polysemy in the academic six-volume monolingual dictionary of Estonian (1st ed. 1988–2007, 2nd ed. 2009), and the supportive theory of generative lexicon by means of a qualia structure (Pustejovsky 1995). The sample of study consisted of simple nouns (843 headwords in all), the total of 1738 semantic units covered both the numbered senses and various subsenses. A hierarchy of the semantic types of nouns, adapted from the lexicographic projects SIMPLE[1] and CoreLex[2], as well as the Estonian Wordnet[3], was used as an ancillary means of analysis enabling, in a way, to "measure" the regularity of alternating word senses.

A result of the analysis is the list of 40 systematically polysemous patterns, presented as the "golden standard" of systematic polysemy in Estonian (named after Peters 2004). In total the sample (843 headwords) contained 305 sense alternations that could be interpreted as revealing systematic polysemy. Of those, nearly every fourth (72 patterns) involves an ARTEFACT sense, while half (!) of the patterns involve ACTIVITY.

1. Introduction

The focus of the study of the polysemy of the Estonian noun (Langemets 2010) was on identifying the systematic patterns of noun polysemy with further perspective to elaborate the principles encode and represent systematic polysemy of nouns in the database of the one-volume dictionary of Estonian (to appear in 2015) and in the **ƎELex** (= EELex) dictionary management system of the Institute of the Estonian Language (Langemets et al. 2006). Systematic pattern encoding enables one to systematize and unify the way semantic information is represented in such language resources, to demonstrate and explain the logical and regular semantic relations between word senses. A systematic list of polysemous patterns could also be of use for automatic semantic analysis of Estonian, on the condition the principle of underspecification (Pustejovsky 1998, Buitelaar 2000) is followed.

2. Methods and material

The analysis was based on the lexical perspective, i.e. on the lexicographic representation of polysemy in the academic six-volume monolingual dictionary of Estonian (DDSE, 1st ed. 1988–2007, 2nd ed. 2009), and the supportive theory of generative lexicon (Pustejovsky 1995). A hierarchy of the semantic types of nouns, adapted from the lexicographic projects SIMPLE[4] and CoreLex[5], as well as the Estonian Wordnet[6], was

[1] SIMPLE homepage, see http://www.ub.es/gilcub/SIMPLE/simple.html (access date 31.03.2010).

[2] CoreLex Online, see http://www.cs.brandeis.edu/~paulb/CoreLex/corelex.html (access date 31.03.2010).

[3] Estonian Wordnet, see http://www.cl.ut.ee/ressursid/teksaurus/ (access date 31.03.2010).

[4] SIMPLE homepage, see http://www.ub.es/gilcub/SIMPLE/simple.html (access date 31.03.2010).

[5] CoreLex Online, see http://www.cs.brandeis.edu/~paulb/CoreLex/corelex.html (access date 31.03.2010).

used as an ancillary means of analysis enabling, in a way, to ‘measure’ the regularity of alternating word senses.

The sample of study consisted of simple nouns, i.e. stem words and derivatives (843 headwords in all, see Table 1). The whole information considered relevant was encoded (manually) to fit the Excel database: the total of 1738 semantic units covered both the numbered senses and various subsenses.

The systematic patterns were analyzed pairwisely by means of a qualia structure (Pustejovsky 1995, SIMPLE), based on the FORMAL ROLES of the nouns, treating the polysemy of nouns denoting artefacts mainly, but other (systematic) combinations have also been analysed. The noun denoting artefact is, according to our definition, any noun with at least one (sub)sense that has ARTEFACT as the FORMAL ROLE.

3. Polysemy of the Estonian noun

Polysemy applies to practically every tenth Estonian word included in the DDSE: 14,432 (10.5%) of the 137,767 headwords have more than one numbered sense (see Table 1). (For the study, a polysemous word has been defined as a DDSE headword supplied with two or more numbered senses.) The most polysemous verbs are *käima* (‘walk’) and *tõmbama* (‘pull’) – with 22 numbered senses both. The most polysemous particle is probably *peale* (‘on, upon, at’) with 21 senses as a postposition, 4 senses as a preposition, and 14 as an adverb (altogether 39 (!) core senses). The most polysemous adjective is *raske* (‘heavy, difficult’) – altogether 18 senses. The most polysemous noun in DDSE is *põhi* (‘bottom’) – altogether 11 senses –, together with the homonymous *põhi* ('north’) with its 4 senses. 11 senses apply also to the noun *vorm* (‘form’), 10 senses to the nouns *ots* (‘head, end’), *pesa* (‘nest’), *vari* (‘shade, shadow’) and the deverbal *käik* (‘gear, walk, stroke’). One can notice the utmost brevity of the most polysemous nouns – 3-4 symbols each –, asserting the Zipf's (1949) law of the least effort. The most polysemous word in the sample is as short: homonym *tee*_H1 (‘road’, 7 senses).

Of Estonian nouns (in the sample, see Table 1), 22% of simple nouns are polysemous. The rest of simple nouns (658 words, 78%) are all monosemous, i.e. with one sense (and thus no special sense number). Of all words, an absolute majority (95%) of compound nouns are monosemous, i.e. polysemy applies to every tenth (10.5%) Estonian word included in the DDSE. Nouns make up about two-thirds of the Estonian vocabulary: our sample estimate is 70%, while in the Estonian Wordnet the noun percentage is 66%.

	Simple nouns	**Compounds (nouns)**	**All words (= DDSE)**
headwords	843	3 200 (provisional data)	137 767
incl. polysemous words	185 (22%)	195 (5%)	14 432 (10.5%)
incl. monosemous words	658 (78%)	3 705 (95%)	123 335 (89.5%)

Table 1. Overview of the sample (shaded in grey)

[6] Estonian Wordnet, see http://www.cl.ut.ee/ressursid/teksaurus/ (access date 31.03.2010).

4. Systematic polysemy of nouns in Estonian

4.1. Nouns denoting artefacts

The author's findings about polysemy of nouns denoting artefacts have been summarized as follows:

(1) most of the interpretations fall in the ARTEFACT class: every second simple polysemous word in the sample (language?) has at least one ARTEFACT-sense (92 of the 185 polysemous words in the sample are nouns denoting artefacts). Of the total of 1738 semantic units, nearly every fifth (315 words) has an ARTEFACT-sense;

(2) from the morphological point of view most nouns denoting artefacts (two-thirds) are stem words, while one third are derivatives;

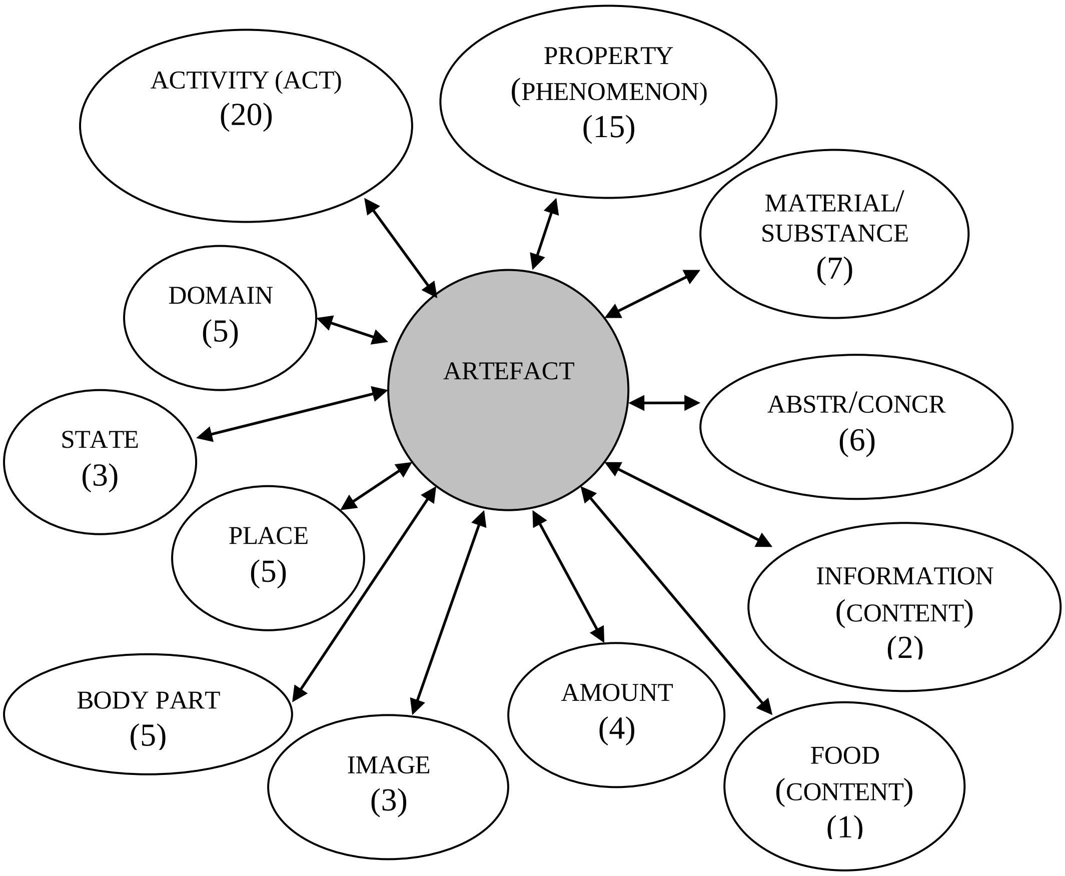

Figure 1. Systematic meaning shifts of ARTEFACT

(3) most of the ARTEFACT-interpretations occupy the position of the first core sense; this is only natural, assuming that an artefact word travels together with its referent. The occurrence of ARTEFACT as a non-first core sense is nearly three times less frequent; e.g. *tafting* qualifies, first and foremost, as an ACT 'method of producing textiles', and only after that as an ARTEFACT 'carpeting' (pattern ACTIVITY–ARTEFACT);

(4) an ARTEFACT-sense usually shares the word with other FORMAL ROLES (semantic types), while nouns standing for ARTEFACT only, without extending to other categories, are relatively rare (20 words in all). The senses of nouns denoting artefacts tend to move on to other categories quite easily, covering a wide range of those. Goddard (1998: 228) has also

emphasized the abundant semantic detail of nouns denoting artefacts, due to which he considers artefact nouns much more complicated semantically than abstract ones;

(5) an ARTEFACT-sense quite often alternates according to the pattern of (metonymic) systematic polysemy, which means that ARTEFACT is quite a generative semantic type (see Figure 1). Of the120 artefact-related semantic pairs in the original list, more than a half, almost two-thirds even (72 pairs) revealed (metonymic) systematic polysemy. Nouns denoting artefacts participate in 12 patterns of systematic polysemy; the three most salient are: ACTIVITY–ARTEFACT (20 cases), e.g. *tikandus* 'obs. embroidery; embroidering', ARTEFACT–PROPERTY (PHENOMENON–ARTEFACT) (15 cases), e.g. *tenor* 'voice; singer', MATERIAL/SUBSTANCE–ARTEFACT (7 cases), e.g. *teras* 'steel; steely quality';

(6) metaphoric polysemy is relatively rare in nouns denoting artefacts. It looks as if an artefact in its essential concreteness hardly inspires figurative creativity and thus ARTEFACT rather becomes the main vehicle of metonymic polysemy. As expected, metaphoric polysemy is more characteristic of more abstract roles such as STATE and PROPERTY, particularly the latter, but also of ACTIVITY and PHENOMENON;

(7) although the first, hypothetical semantic structure for ARTEFACT mainly consisted of the AGENTIVE and TELIC ROLES, upon analysis the CONSTITUTIVE ROLE has been added, because in more than half the cases systematic polysemy in artefact word patterns has been interpreted by the CONSTITUTIVE ROLE (39 cases):
ARTEFACT = [CONCRETE_ENTITY : ARTEFACT | AGENTIVE ROLE | TELIC ROLE | CONSTITUTIVE ROLE]

4.2. Nouns denoting other semantic types

Systematic polysemy of the other semantic types, i.e. the partner elements of ARTEFACT, is slightly less extensive than that of ARTEFACT words.

The second biggest bundle of polysemous patterns (10 different systematic patterns) after ARTEFACT gravitates around PROPERTY—the content and character of an entity, like a name for the contents of a vessel, serving at the same time as a name for the vessel itself, be it either ARTEFACT, PLACE, HUMAN (or ANIMAL), STATE, PHENOMENON, ACT (incl. SPEECH ACT) or AMOUNT, or some other entity possibly represented by a single noun in our sample. In semantic alternation, PROPERTY shows considerable preference for PHENOMENON, ARTEFACT and HUMAN.

Another large bundle (9 systematic patterns) is made up by ACTIVITY most frequently participating in semantic alternation with PROPERTY, ARTEFACT and ACT/SPEECH ACT, and also with ABSTRACTION/CONCRETE_ENTITY and STATE. In addition there are EVENT, PLACE and/or INSTITUTION/BUILDING, DOMAIN and AMOUNT. For different words those components co-occur in different combinations.

The interpretation of HUMAN may be represented in 7 systematic patterns, involving, first and foremost, nouns of PROPERTY and REPRESENTATION (partly close to the former), and also ACTIVITY_AGENT and PLACE, as well as INSTITUTION. Alternation of part–whole in BODY PART is also typical.

5. The gold standard of systematic polysemy patterns in Estonian

A result of the analysis is the list of 40 systematically polysemous patterns, presented as the ‘golden standard’ of systematic polysemy in Estonian (named after Peters 2004).

The number of detected and selected regular semantic shifts varies across languages, depending on the scope and purpose of the project: in English – 126 underspecified semantic types (Buitelaar 2000), 138 patterns of regular polysemy (Peters 2004: 151–166); in SIMPLE Finnish lexicon – 16 patterns (Salmisuo 2000); in Dutch database and dictionary system – 11 regular meaning shifts (Vliet 2007).

In total the sample (843 headwords) contained 305 sense alternations that could be interpreted as revealing systematic polysemy. Of those, nearly every fourth (72 patterns) involves an ARTEFACT sense, while half (!) of the patterns involve ACTIVITY (shaded grey in Table 2). An absolute majority of systematically polysemous nouns belong to a single pair of regular semantic transfer. A score of nouns capture two different systematic patterns. In the sample, institution nouns (*teater* ‘theatre’) were the richest in semantic transfers, involving many various elements, and so were event and act(ivity) nouns (*tee* ‘tea’, *tants* ‘dance’).

Pattern	**Number of cases**	**Example**
ACTIVITY–ACT/SPEECH ACT	26	*taotlus* ‘application’
ACTIVITY–ABSTR/CONCR	21	*tasu* ‘payment, compensation’
ACTIVITY_AGENT–PROPERTY	21	*tantsija* ‘dancer’
ACTIVITY–ARTEFACT	20	*tikandus* ‘obs. embroidery; embroidering’
PHENOMENON–PROPERTY	16	*teadvus* ‘consciousness’
ARTEFACT–PROPERTY (PHENOMENON–ARTEFACT)	15	*tenor* ‘tenor (instrument; voice)’
PROPERTY–HUMAN (ANIMAL)	14	*tenor* ‘tenor (voice; singer)’
ACTIVITY–STATE	13	*terror* ‘terror’
PLACE/INSTITUTION/BUILDING–ACTIVITY (SOCIAL EVENT)	11	*teater* ‘theater’
ACTIVITY–DOMAIN	9	*tants* ‘dance’
ABSTR/CONCR –STATE	8	*tegelikkus* ‘reality’
ABSTR/CONCR –DOMAIN	8	*tantrism* ‘tantrism’
ABSTR/CONCR –INSTITUTION	7	*telefon* ‘telephone’, *ooper* ‘opera’
MATERIAL/SUBSTANCE–ARTEFACT	7	*tegelikkus* ‘reality’
ABSTR/CONCR–ARTEFACT	6	*teras* ‘steel; steely quality’
PROPERTY–AMOUNT	6	*tihedus* ‘density’
PLANT–FOOD	6	*tee* ‘tea’
ARTEFACT–PLACE	5	*tagala, tara*
ARTEFACT–DOMAIN	5	*joonistus* ‘drawing’
BODY PART–ARTEFACT	5	*talje* ‘waist; vest’
PROPERTY–SPEECH ACT	5	*tarkus* ‘wisdom’
PROPERTY–STATE	5	*taibutus* ‘insensateness’
ACTIVITY–SOCIAL EVENT	5	*teater* ‘theater’

Table 2. The most frequent systematic polysemy patterns of nouns in Estonian

5. Representing systematic polysemy

The research on systematic polysemy should be of help to differentiate contrastive and complementary polysemy – a principle long emphasized and desired in lexical semantics (Weinreich 1964). Complementary polysemy, incl. (a bundle of) systematic polysemy, would be systematically represented as a subsense in the database of Estonian. Of course, it is reasonable to represent *patterns currently in use*, maybe even just the *more frequent* ones (see Van der Vliet 2007). If a subsense is of a different kind, for example metaphorically polysemous, it should appear under a different marker (e.g. *fig*.).

References

Buitelaar, P. (2000). ‘Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification’. In *Proceedings of the ANLP2000: Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems*. Seattle, USA. http://dfki.de/~paulb/anlp00.html (access date 30.03.2010).

DDSE (1988–2007) = Karelson, R.; Kullus (Põlma), V.; Raiet, E.; Tiits, M.; Valdre, T.; Veskis, L. (eds.). *Eesti kirjakeele seletussõnaraamat* (26 vihikut). [= *Defining Dictionary of Standard Estonian* (vehicles 1–26)]. Tallinn: Eesti Keele Sihtasutus.

DDSE (2009) = Langemets, M.; Tiits, M.; Valdre, T.; Veskis, L.; Viks, Ü.; Voll, P. (eds.). *Eesti keele seletav sõnaraamat* I–VI [= *Defining Dictionary of Estonian* I–VI]. 2nd ed. Tallinn: Eesti Keele Sihtasutus, 2009.

Goddard, C. (1998). *Semantic Analysis: A Practical Introduction*. Oxford: Oxford University Press.

Langemets, M. (2010). *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras* [= *Systematic Polysemy of Nouns and its Lexicographic Treatment in Estonian*]. Tallinn: Eesti Keele Sihtasutus.

Langemets, M.; Loopmann, A.; Viks, Ü. (2006). ‘The IEL Dictionary Management System of Estonian’. In de Schryver, G.-M. (ed.). *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems*. Turin: Turin University. 11–16.

Peters, W. (2004). *Detection and Characterization of Figurative Language Use in WordNet*. PhD thesis. Natural Language Processing Group, Department of Computer Science, University of Sheffield. http://www.dcs.shef.ac.uk/~wim/thesis_web.pdf (access date 08.01.2010).

Pustejovsky, J. (1995). *The Generative Lexicon*. The MIT Press.

Pustejovsky, J. (1998). ‘The semantics of lexical underspecification’. In *Folia Linguistica* 32. 323–347.

Salmisuo, Sari 2000. *SIMPLE lexicon documentation* [Finnish lexicon]: http://www.ub.es/gilcub/SIMPLE/reports/simple/D35_HELrev.htm (access date 08.01.2010).

SIMPLE = Lenci, A.; Busa, F.; Ruimy, N.; Gola, E.; Monachini, M.; Calzolari, N.; Zampolli, A. (2000). *Linguistic Specifications. SIMPLE Work Package 2. Deliverable D2.1*.

Van der Vliet, H. (2007). ‘The Referentiebestand Nederlands as a multi-purpose lexical database’. In *International Journal of Lexicography* 20 (3). 239–258.

Weinreich, U. (1964). ‘Webster's Third: A Critique of its Semantics’. In *International Journal of American Linguistics* 30 (4). 405–409. [Reprint in William Labov, W.; Weinreich, B. S. (eds.) (1980). *On Semantics*. Philadelphia: University of Pennsylvania Press. 361–367.]