LexikoNet - a lexical database based on type and role hierarchies

Alexander Geyken*, **Norbert Schrader***

*Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstr. 22/23, 10117 Berlin, www.dwds.de
geyken@bbaw.de
{geyken, schrader}@bbaw.de

Abstract

In this paper LexikoNet, a large lexical ontology of German nouns is presented. Unlike GermaNet and the Princeton WordNet, LexikoNet has distinguished type and role hypernyms right from the outset and organizes those lexemes in a parallel, independent hierarchy. In addition to roles and types, LexikoNet uses meronymic and holonymic relations as well as the instance relation. LexikoNet is based on a conceptual hierarchy of currently 1,470 classes to which approximately 90,000 word senses taken from a large German monolingual dictionary, the Wörterbuch der deutschen Gegenwartssprache (WDG), are attached. The conceptual classes provide a useful degree of abstraction for the lexicographic description of selectional restrictions, thus making LexikoNet a useful filtering tool for corpus based lexicographic analysis. LexikoNet is currently used in-house as a filter for lexicographic extraction tasks in the DWDS project. Furthermore, it is used as an classification tool of the 'words of the week' provided for the newspaper *Die ZEIT* on www.zeit.de.

1. Introduction

LexikoNet is a large lexical ontology of German nouns developed by Alexander Geyken and Norbert Schrader (Berlin-Brandenburg Academy of Sciences - BBAW). LexikoNet was developed as a corpus filtering tool for the lexicographic analysis of selectional restrictions in the context of a German dictionary project, the Digital Dictionary of the German language (DWDS) at the BBAW (Klein and Geyken, 2004; Geyken, 2006).

The development of LexikoNet started in 2000 with three motivations. First, there was no freely available, sufficiently large lexical ontology of nouns for the everyday German language: Neither GermaNet, the German WordNet (Kunze, 2000), nor the semantic hierarchy of CISLEX (Langer, 1996) were freely available. Hence, it would have been difficult to integrate those ontologies into the linguistic search engine of the DWDS website (www.dwds.de). Furthermore, none of these ontologies applies the role-type distinction systematically but they rely on a single hypernym relation between subordinate and superordinate concepts. However, researchers in the field of knowledge representation examining the semantics of the relations in semantic networks have distinguished different types of is_a relations for more than 30 years. One of the most influential papers has been Wood's paper on a typology of is_a links (Woods, 1975). A few years later, semantic networks like Brachman's KL-One used the difference between type and role for the semantic representation of terminological knowledge (T-Box) (Brachman, 1983). Therefore, it seems intuitively clear that this distinction should also be applied to lexical ontologies covering general language. The third reason for the creation of a new lexical ontology was that often the synset itself is not the appropriate level of abstraction for the lexicographic description of selectional restrictions. The Princeton WordNet, for example, introduces non-lexicalized concepts such as 'bad person' (characterized as a person who does harm to others) or 'adult female' (characterized as an adult female person (as opposed to a man)) only for perceived lexical gaps. In LexikoNet, we base the entire description on a conceptual hierarchy that is generally not lexicalized. In particular, concepts referring to the role relation are frequently expressed by phrases. For example, the role relation of 'donkey' above is expressed by a complex noun phrase DOMESTIC BEAST OF BURDEN. Similarly, many human nouns can be subclassified according to their conceptual attributes. Persons can be classified according to their physical properties or their belief. In many cases, for a given lexeme only one attribute is predominant. For example, lexemes like 'Christian' or 'Moslem' can be classified as PERSONS BY RELIGIOUS BELIEF, 'adonis' or 'dwarf' as PERSONS BY PHYSICAL CHARACTERISTICS, or 'candidate' or 'nominee' as PERSONS BY SOCIAL CONTEXT. We will show in section 4 how these concepts apply in the description of selectional restrictions for the different meanings of the verb 'aufstellen' (engl. nominate, put up, compose).

2. Description of LexikoNet

LexikoNet is based on a concept hierarchy of currently 1,470 concept nodes that is ordered in a top-down hierarchy beginning with the concepts of CONCRETE NOUNS (1186 nodes) and ABSTRACT NOUNS (284). CONCRETE NOUNS are further subdivided into LIVING BEINGS, PHYSICAL OBJECTS, SUBSTANCES AND MATERIALS, and SPACE AND PLACES. ABSTRACT NOUNS subdivide into PROPERTY, EVENT, ACTIVITY, MEASURES, DOMAINS, etc. The concept hierarchy of ARTIFACTS (underneath PHYSICAL OBJECTS) or LIVING BEINGS, for example, subdivide some levels deeper into rather specific categories such as SACRAL BUILDING or SPORTS TEAM. The hierarchy goes up to 10 levels deep. The following list gives an overview of the conceptual hierarchy and displays up to four levels where each indentation corresponds to one level:

concrete nouns
→ living beings
  → hominids

- → human species, e.g. Neanderthal man
- → human being, e.g. individual, sorry mortal
  - → person by name, e.g. Angela, Blair, PhD
  - → person by characteristics
    - → person by physical characteristics e.g. woman, old man, athlete
    - → person by mental characteristics e.g. genius, lunatic, optimist
    - → person by social characteristics e.g. father, Asian, communist
    - → person by activity e.g. accomplice, bank employee, patient
    - → person by outfit or equipment e.g. gunman
- → social unit
  - → social category e.g. mankind, younger generation, society
  - → crowd or gathering, e.g. mob, queue
  - → couple, e.g. married couple
  - → primary group, e.g. family, clique
  - → specific administration union e.g. sports club, company, government
  - → social unit by activity e.g. following (followers)
- → anthropomorphous supernatural being e.g. goddess, devil
- → animals
  - → animal by biological classification e.g. mammal, humming bird
  - → animal by natural characteristics e.g. female, ruminant
  - → animal by practical characteristics e.g. beast of burden, ornamental fish
  - → unit of several animals e.g. fauna, colony of bees
- → plants and mushrooms
  - → plant
    - → plant by biological classification e.g. conifer, horsetail
    - → plant by typical form e.g. tree, herb
    - → plant by natural characteristics e.g. seedling, epiphyte
    - → plant by practical characteristics e.g. winter barley, weed
    - → unit of several plants, e.g. grove
  - → mushroom, e.g. boletus, dry rot
  - → lichen, e.g. foliose lichen
- → microorganisms and virus e.g. amoeba, bacteriophage
- → taxonomical unit, e.g. species, order
- → physical objects
  - → natural thing e.g. body, leg, trunk, pebble, cloud, molecule
  - → artifact e.g. hammer, aircraft, saxophone, skyscraper
- → substances and materials
  - → material by origin e.g. wood, metal, water, air
  - → material by function e.g. nutrient, fuel, waste
  - → chemical element or compound e.g. helium, iron oxide
  - → material by aggregate state e.g. ice, liquid, steam
  - → quantity of material, e.g. clod
  - → mythological material e.g. philosopher's stone
- → space and places
  - → space by nature e.g. atmosphere, ocean, Africa, lowlands
  - → space by use e.g. housing area, field, highway
  - → space by possession e.g. leasehold land
  - → place by characteristics of form e.g. edge, surface, cavity
  - → mythological place, e.g. paradise
- → geometric configurations e.g. curve, triangle, ellipsoid

abstract nouns

- → abstract spaces, e.g. infinite space, nowhere
- → abstract objects, e.g. object, part, nothingness
- → qualities properties and state e.g. color, intelligence
- → events, e.g. tsunami, Olympic Games
- → activities and behavior, e.g. work, jogging, laughter
- → patterns of activity, e.g. method, language
- → ideas and information e.g. concept, theory of relativity, party platform
- → domains and disciplines, e.g. geography, handicraft
- → cultures and social systems e.g. Mayan culture, democracy
- → time and history, e.g. season, evening, Middle Ages
- → numbers and measures, e.g. hundred, gallon

Currently, some 90,000 lexemes (corresponding to 75,000 different types) are associated to the concept hierarchy. Lexemes in the sense of LexikoNet are dictionary senses taken from a large German monolingual dictionary, the Wörterbuch der deutschen Gegenwartssprache (WDG, (Klappenbach and Steinitz, 1977)). Two dictionary senses are mapped onto one sense if they are not distinguished in the concept hierarchy. Lexemes are associated to concept nodes at the most specific level by three kinds of IS-A relations (type, role and instance) as well as a meronymic and a holonymic relation.

Instances and generic terms are separated in order to enable a systematic term enrichment. For example, LexikoNet distinguishes the concepts SACRAL BUILDING and SACRAL BUILDING BY NAME. The former contains terms like 'cathedral', 'synagogue', 'mosque' etc., the latter terms named entities like 'Stephansdom', 'Paulskirche' or 'Al-Aqsa mosque'. In the same way 'football team' or 'basketball team' (in German those nouns are compounds) are subsumed by the concept SPORTS TEAM, whereas 'Arsenal London' or 'Inter Mailand' are classified as instances of the concept SPORTS TEAM BY NAME.

The difference between type and role is important in many semantic fields such as animals, artifacts, plants, professions, or events which all can be classified according to their generic resp. biologic classification or with respect to their (anthropocentric) function. For example, the abovementioned 'mosque' is a building having a sacral role or a person can have the role as a father or a friend. In many cases, the same lexeme can have both, a type and a role. For example, the lexeme 'donkey' in the type hierarchy is a kind of ODD-TOED UNGULATE whereas in the role hierarchy it is a kind of DOMESTIC BEAST OF BURDEN. Unlike WordNet, GermaNet and CISLEX, LexikoNet systematically organizes those lexemes in a parallel, independent hierarchy, i.e. a type and a role hierarchy. Formally speaking, this means that the hierarchy of LexikoNet corresponds to a lattice and not to a tree.

Approximately 30,000 lexemes in LexikoNet are simple base forms containing exactly one stem. The rest are compounds. Stems in LexikoNet are directly related to the TAGH-morphology, a complete morphology of German which accounts for derivation and composition (Geyken and Hanneforth, 2006). Compounds are included in the LexikoNet if their meaning is not compositional or if the meaning of the suffix component is ambiguous. If we describe a compound as consisting of two components A and B then non-compositionality means that the meaning of compound AB in the conceptual hierarchy of LexikoNet does not correspond to the meaning of B in LexikoNet. For example, a true compound like 'Eisenhut' (engl. wolfsbane) is subsumed under the concept node PLANT which is different from the semantics of `Eisen#Hut` (`iron#hat`) where the B-component 'Hut' is classified as CLOTHING. Compounds are also included in the LexikoNet if the meaning of the B-component is ambiguous with respect to the concept nodes of LexikoNet. For example, a word like 'Lebensversicherung' (life insurance) would be included since the B-component 'Versicherung' (engl. affirmation, assurance or insurance) corresponds to different conceptual nodes in LexikoNet. No new lexeme is added in the remaining case where the meaning of the compound AB and the B-component B are subsumed by the same concept node in LexikoNet. For example, the compounds 'Holztür (wood door) or 'Wohnzimmertür' (living room door) which are decomposable into `Holz#Tür` resp. `Wohnzimmer#Tür` are not part of the LexikoNet since their meaning is inherited from the B-component 'Tür' even though both compounds differ in their qualia roles.

Lexemes in LexikoNet are not only related to concept nodes, they can also be interrelated by role and type relations as well as meronymic and holonymic relations. An example for this is the lexeme 'airport' which has several meronyms such as 'terminal', 'runway' or 'fingerdock'. Those meronyms in turn are related to the concept nodes BUILDING BY ITS ROLE FOR TRANSPORTATION or LOCATION BY ITS ROLE FOR TRANSPORTATION. Moreover, lexemes that are subsumed by the same concept node can be related by a hypernym relation. For example, 'hurricane' is an hyponym of 'cyclone' which in turn is an hyponym of 'windstorm'. However, in LexikoNet they are subsumed under the concept node METEOROLOGICAL PHENOMENON. Here, LexikoNet follows the hierarchy of WordNet. We will see in the next section, how LexikoNet and WordNet differ in the way they consider the hypernym hierarchy.

3. LexikoNet vs. WordNet

Unlike LexikoNet, WordNet does not distinguish roles and types in the hypernym hierarchy. We give some arguments that this difference matters not only with respect to completeness of the lexical database but also for concrete information extraction tasks. Consider the following examples. The word 'donkey' in WordNet (WordNet 2.1) has only the type reading even though, as shown above, it should also have the role reading:

domestic ass, donkey, Equus asinus (domestic beast
of burden descended from the African wild ass)
  ⇒ odd-toed ungulate, perissodactyl,
    perissodactyl mammal (placental mammals having
    hooves with an odd number of toes on each foot)
    ⇒ mammal

A less well known example is the word 'hurricane'. In WordNet 2.1 the following inherited hypernym hierarchy is given:

hurricane
⇒ a severe tropical cyclone usually
  with heavy rains and winds
  ⇒ cyclone (a violent rotating windstorm)
    ⇒ windstorm (a storm consisting of violent winds)
      ⇒ storm, violent storm (a violent weather
        condition with winds (11 on the Beaufort scale)
        and precipitation and thunder and lightening)
        ⇒ atmospheric phenomenon (a physical
          phenomenon associated with the atmosphere)

Here, WordNet encodes the type hierarchy. However the role hierarchy is not encoded. Yet it might be very useful for information extraction applications to know that 'hurricanes' are events that often (note here that roles are defeasible) are related to a natural catastrophe. This double classification should also be done for the term 'tsunami' which in WordNet (2.1) is only classified according to its - admittedly - predominant function (role), but not to its type as a natural phenomenon:

tsunami (a cataclysm resulting from a destructive
sea wave caused by an earthquake or volcanic eruption)
  ⇒ calamity, catastrophe, disaster, tragedy
  cataclysm (an event resulting in great loss and misfortune)

4. LexikoNet and selectional restrictions

The conceptual nodes of LexikoNet provide a useful level of abstraction for the description of selectional restrictions. LexikoNet intended here as a filter for corpus queries. We additionally presuppose that the corpus is annotated with syntactic functions by a shallow parser. It is then possible to extract all direct objects of a given verb and to classify those with LexikoNet's concept nodes. The resulting concept bundle will then provide a more appropriate view about the word senses than by a pure word based approach. We will illustrate this by the German verb 'aufstellen' which has six senses in the WDG. Because of space limitations, we will focus here on the first three senses:

'Aufstellen', sense 1, means 'to arrange' or 'to assemble'. In both subsenses, 'aufstellen' selects artifacts for its direct object. For lexicographic purposes this must be refined as 'etwas Größeres aus Teilen an einem Platz errichten' (to assemble something larger out of its component parts) or 'etwas an einem Platz in eine stehende Position bringen' (to bring something into a standing position). Both senses correlate with LexikoNet's concept nodes BEHELFSUNTERKUNFT (provisional accomodation), ERHOEHTE PLATTFORM (scaffolding) or UNTERGESTELL (undercarriage).

'Aufstellen', sense 2, means 'to nominate'. Here, 'aufstellen' selects 'humans in the social context of applying to a position or an award'. This corresponds to a small class of MENSCH ALS BEWERBER (human as candidate), a subclass of the concept MENSCH IM SOZIALEN HANDLUNGSKONTEXT (person in a social context), which includes lexemes like 'Kandidat' (candidate), 'Bewerber' or 'Postulant' (both applicant). Note here also the difference with WordNet 2.1 where both synsets do not provide a conceptual generalization since they are both directly subsumed under 'person':

candidate, prospect (someone who is considered for something (for an office or prize or honor etc.))
⇒ person, individual, someone, somebody
applicant, applier (a person who requests or seeks something such as assistance or employment or admission)
⇒ person, individual, someone, somebody

'Aufstellen', sense 3, means 'to set up' like in *to set up a sports team*, or 'to deploy' like in *to deploy a military unit*. LexikoNet provides here the conceptual nodes ZWECKORIENTIERTE KLEINGRUPPE (goal-oriented small group) and MILITAERISCHE EINHEIT (military unit). Here again, it would be difficult to find the appropriate synset level in WordNet. The synset 'crowd' is too general with respect to 'goal-oriented small group' whereas the synsets 'army' or 'military' are too specific with respect to 'military unit' (which by the way exists as a head noun in the glosses of the synset 'army').

5. Applications and further work

LexikoNet is currently used in-house as a filter for lexicographic extraction tasks in the DWDS project. LexikoNet is also applied in the classification of frequently occurring nouns, the so-called *words of the week* in the newspaper *Die ZEIT* (www.zeit.de). In this application, only a very small subset of LexikoNet is used: 'persons', 'organizations', 'plants', 'animals', 'events', 'geographical nouns', 'material', 'food' and 'diseases'. Here we find an obvious practical use for distinguishing a word's type and role functions. For example, the word 'hurricane' in a newspaper context is more likely to occur in its role relation, i.e. as a 'natural catastrophe', than in its type relation, i.e. as a 'atmospheric phenomenon'. Hence it should preferably be classified as an 'event' in the context of the *words of the week*. This has been implemented by a precedence rule 'event' ≫ 'atmospheric phenomenon', meaning that the 'event' reading of a lexeme is the preferred over its reading as a 'atmospheric phenomenon'. The result of this classification for the last 100 issues of *Die ZEIT* can be found under www.dwds.de/woewo.

Future work on LexikoNet will focus on term enrichment with additional compounds on the basis of the electronic version of the above-mentioned monolingual WDG dictionary. LexikoNet is currently undergoing a correction phase. A first public release is planned for early 2007.

6. References

Ronald J. Brachman. 1983. What is-a is and isn't: An Analysis of Taxonomic Links in Semantic Networks. *IEEE Computer*, 16 (10).

Alexander Geyken and Thomas Hanneforth. 2006. TAGH: A Complete Morphology for German based on Weighted Finite State Automata. *Proceedings of FSMNLP 2005, Lecture Notes in Artificial Intelligence*.

Alexander Geyken. 2006. DWDS-Corpus. In Christiane Fellbaum, editor, *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*. Continuum Press, London.

Ruth Klappenbach and Wolfgang Steinitz, editors. 1977. *Wörterbuch der deutschen Gegenwartssprache*, volume 1–6. Akademie-Verlag, 1st edition.

Wolfgang Klein and Alexander Geyken. 2004. Projekt Digitales Wörterbuch der deutschen Sprache des 20. Jh. In *Jahrbuch der BBAW 2003*. Akademie Verlag, Berlin.

Claudia Kunze. 2000. Extension and Use of GermaNet, a lexical-semantic database. *Proceedings of the Second International Conference on Language Resources and Evaluation*, II:999–1002.

Stefan Langer. 1996. *Selektionsklassen und Hyponymie im Lexikon. Semantische Klassifizierung von Nomina für das elektronische Wörterbuch CISLEX*, volume 94 of *CIS-Bericht*. München.

William Woods. 1975. What's in a link: Foundations for Semantic Networks. In Bobrow D.G. and Collins A.M., editors, *Representation and Understanding: Studies in Cognitive Science*, pages 35–82. Academic Press, New York.