

Semi-automatic Extraction and Classification of Predicates from German Text Corpora

Ekaterina Lapshinova-Koltunski
katerina@ims.uni-stuttgart.de
IMS, University of Stuttgart

March 13, 2008
AAACL-2008, BYU, Provo

Examples of nominal predicates

1. *Allein **die Ankündigung**, dass er komme, hatte den Börsenkurs vergangene Wochen in die Höhe getrieben.* (“Alone **the notification**, that he would come, boosted the market rate in the bygone weeks”.)
2. *Selbst **die glasklare Erfahrung**, dass der erste Verlust immer noch am wenigsten schmerzt, fehlt ihnen zumeist.* (“Even **the crystal clear experience** that the first loss hurts still lost of all, is mostly missing”)

Examples of multiword predicates

(a)+(b)

1. *Darauf gingen Rektoren und Präsidenten um so bereitwilliger ein, als ihnen **in Aussicht gestellt wurde**, dass die Hochschulleitungen gestärkt werden sollten...* (“Rectors and presidents were all the more interested, as it was **put into perspective (announced)** that the university administration should be increased...”)
2. *Weil die Lieferländer aber **zur Bedingung gemacht haben**, dass die Waren zu freien Marktpreisen verkauft werden...* (“Because the supplier countries **made it a condition** that the goods should be sold for free market prices.”)
3. *...daß Ihr Beauties immer dann am unterhaltsamsten seid, wenn Ihr krampfhaft **unter Beweis stellen wollt**, dass in Deutschland seit neuestem auch Mädels Abi machen dürfen.* (“...that you beauties are always most amusing when you **are giving proof of** that in Germany just since very recently also gals are allowed to do the high school degree.”)
4. *Weil Clinton und seine Anwälte erst **in Erfahrung bringen wollen**, was Lewinsky zu sagen hat.* (“Because Clinton and his lawyers **into experience to bring want (want to find out)**, what Lewinsky has to say.”)
5. *Wenn **in Rechnung gestellt wird**, dass Mädchen und Jungen auf ungleiche Weise lernen, aber zu gleichwertigen Lernergebnissen gelangen...* (“**Bringing to account** that girls and boys learn in a different way, but succeed to equal learning results...”)

(c)+(d)

6. *Ich will nicht **in Abrede stellen**, dass in diesem Bereich in Deutschland durchaus noch Anstrengungen notwendig sind.* (“I don’t want to **put into understanding (to deny)** that more effort are absolutely required in this field in Germany.”)

7. *wenn dem Wessi **ins Auge fällt**, dass in der Hauptstadt der Liter Dieselkraftstoff für 1,12 statt eine Mark wie in Stuttgart gezapft wird...* (“when it **catches the “Wessi’s”(West German’s) eye** that a liter of diesel fuel in the capital is sold for 1,12 and not just one mark as in Stuttgart...”)
8. *...Womit wohl überdeutlich **zum Ausdruck kam**, dass nicht alle Tassen im Schrank waren.* (“...Whereby it obviously **came to expression (was expressed)** that they don’t have both oars in the water.”)
9. *...wobei allerdings nicht **zur Sprache kam**, dass sich Berlusconi selbst mit der British Telecom verbündet hat.* (“...whereby it **didn’t come to speech (wasn’t expressed)** that Berlusconi allied himself with British Telecom. ”)
10. *...damit nicht ganz **in Vergessenheit gerate**, dass auch hinter den Fenstern eine Welt liegt.* (“...so that it is not fully **buried in oblivion** that there is a world also behind the windows.”)

Examples of nominal compound predicates

(a)

1. *Die Journalisten**frage**, ob es denn kompromittierende Fotos von ihm gebe...* (“The journalist **question** if there are compromising photos of him...”)

(b)

2. *Aber die **Erklärungsversuche**, warum der Teufel sich an die Frau Doktor heranmacht, sind auf der Glatze gedrehte Locken.* (*‘DEREKO-AT’*) (“But all the **expansion**-attempts (attempts to explain) why the devil chats up the female doctor are as futile as giving a bald man a comb”)
3. *Das **Auswahlverfahren**, wer Christine Frisinghelli ab dem Jahr 2000 an der Spitze des “steirischen herbstes” nachfolgen wird...* (“The **selection** process, who will follow Christine Frisinghelli at the top of the ‘steirischer herbst’ festival as of 2000...”)
4. *Als **Beweismittel**, dass die Frauen alles freiwillig gemacht hätten, legte der Verteidiger dem Gericht ein rotes Dessous eines angeblichen Opfers vor.* (“As **means of evidence** that women had done everything voluntary the defense lawyer presented the red underwear of the alleged victim to the court.”)
5. *Der **Kontrollgriff**, ob noch alles da ist, nervt spätestens am dritten Tag.* (“The **control** grip, if everything is still there, nerves at the latest on the third day.”)
6. *Die einst gewünschte **Bedenkzeit**, ob er das Amt annehmen solle..., dünkt ihm im Rückblick jedenfalls viel zu lang....* (“The once desired **consider time (time of consideration)**, if he should accept office..., seems to have been too long for him...”)
7. *Das **Erläuterungsbemühen** des Wissenschaftsrates, wer aus welchen Gründen in die Evaluierungskommission berufen wurde, hielt sich bisher in engsten Grenzen.* (“The **explanation** effort of the science council, who and for what reasons was appointed to the evaluation commission, was kept so far within a limit.”)

(c)

8. *In dem edlen Bonner **Wettstreit**, wer die Kunst des Nebelwerfens und Schaumschlagens am besten beherrscht, rückt ein Mann unaufhaltsam nach vorn: Rudolf Scharping.* (“In the noble Bonn **competition**, who is the best smoke launcher and hot-air merchant, one man moves forward inexorably: Rudolf Scharping”.)

9. *Das **Rätse**lraten, wer in der Regierung sitzen soll, geht damit los.* (“The **riddle solution (guess-work)** who should sit in the government goes off with it.”)
10. *Die sportliche **Wunsch**vorstellung, dass die Besten der winterlichen Wettkampfsreihe auch beim punktuellen Großereignis als Sieger gefeiert werden, ist keine logische Konsequenz.* (“The sport **desire association (wishful thinking)** that the best of winter competition series will also be celebrated as winner at the selective huge event is not consequential.”)
11. *Den **Ehrgeiz**, dass sie die jungen Menschen zu "wetterfesten Persönlichkeiten" erziehen, erwartet der Bundespräsident von allen Lehrern und von allen Schulen.* (“The **honour avarice (ambition)** that they bring up young people as 'weatherproof characters', the president expects from all the teachers and schools.”)
12. *Den **Volk**smund, dass Kleinvieh auch viel Mist machen kann, bestätigen die Fluggesellschaften mit ihren Vielfliegerprogrammen.* (“The **common parlance** that many a little makes a mickle is confirmed by airlines with their frequent flyer programs.”)

References

- [Butt et al. 2002] M. Butt, H. Dyvik, T.King, H. Masuichi, C. Rohrer: “The Parallel Grammar Project”. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pp. 1-7.
- [Copestake et al. 2004] A. Copestake, F. Lambeau, B. Waldron, F. Bond, D. Ickinger, S. Oepen: “A lexicon module for a grammar development environment”, in: *Proceedings of the Linguistic Resources and Evaluation Conference 2004*, Lisboa, Portugal, 2004, pp. 1111-1114
- [Eckle-Köhler 1999] Eckle-Köhler, J. (1999), *Linguistic Knowledge for Automatic Lexicon Acquisition from German Text Corpora*. Berlin: Logos Verlag.
- [ELDIT] <http://dev.eurac.edu:8081/MakeEldit1/Eldit.html>
- [Evert 2005] Evert E. (2005). The CQP Query Language Tutorial. IMS, Stuttgart. URL <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/html/>
- [Fellbaum et al.] Fellbaum C., A. Geyken, A. Herold, F. Koerner, and G. Neumann (2006) . Corpus-based studies of German idioms and light verbs, in *International Journal of Lexicography*, vol. 19:4.
- [Grishman et al. 1994] Grishman, R., C. Macleod and A. Meyers (1994). “COMLEX Syntax: Building a Computational Lexicon”, Presented at Coling 1994, Kyoto.
- [Gurevich et al. 2007] Gurevich, O., R. Crouch, T.H. King, V. de Paiva (2007). Deverbal Nouns in Knowledge Representation. In *Journal of Logic and Computation Advance Access*. December 20.
- [Heid/Gows 2006] Heid, U. and R.Gows (2006). A model for a multifunctional electronic dictionary of collocations. In *Proceedings of the XIIth Euralex International Congress*, Torino: pp. 979-988.
- [Helbig/Buscha 2005] Helbig, G., J.Buscha (2005). *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht*. Berlin, Langenscheidt.
- [Herbst et al. 2004] Herbst, T., D. Heath, I.F. Roe and D.Götz (2004). *A Valency Dictionary of English. A Corpus-Based Analysis of English Verbs, Nouns and Adjectives*. Berlin/New York: Mouton de Gruyter.
- [Kermes 2003] Kermes, H. (2003). *Off-line (and On-line) Text Analysis for Computational Lexicography*. Ph.D. thesis IMS, University of Stuttgart. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)*, volume 9, number 3.
- [Lapshinova 2007] Lapshinova, E. (2007). *Extracting Predicates Subcategorizing for Wh-Clauses: an Architecture for a Semi-automatic System*. In *Proceedings of the 12th ESSLLI Student Session*. Dublin, Ireland, August 6-17, 2007.

- [Lapshinova 2008] Lapshinova, E. (2008). Non-heads of compounds as valency bearers: extraction from corpora, classification and implication for dictionaries. In *Proceedings of EURALEX-2008*, Barcelona. Spain.
- [Lapshinova/Heid 2007] Lapshinova, E., U.Heid (2007). Syntactic subcategorization of noun+verb multiwords: description, classification and extraction from text corpora. In: *Proceedings of the 26th International Conference on Lexis and Grammar*. Bonifacio, Corse, October 2-6, 2007.
- [Lapshinova/Heid 2008] Lapshinova, E., U.Heid (2008). Head or Non-head? Semi-automatic procedures for extracting and classifying subcategorisation properties of compounds. In *Proceedings of LREC-2008*. Marrakech, Marokko.
- [Lezius/Dipper/Fitschen 2000] Lezius, W., S. Dipper and A. Fitschen (2000). IMSLex - Representing Morphological and Syntactical Information in a Relational Database. In U. Heid, S. Evert, E. Lehmann and C. Rohrer. (Hrsgg.), *Proceedings of EURALEX*, Stuttgart, Germany, pp. 133-139.
- [Macleod et al. 1998] Macleod, C., R. Grishman, A. Meyers, L. Barrett, R. Reeves. NOMLEX: A Lexicon of Nominalizations. *Proceedings of EURALEX'98*, Liege, Belgium, August 1998. <http://nlp.cs.nyu.edu/nomlex/NOMLEX-2001.reg>
- [NOMLEX] <http://nlp.cs.nyu.edu/nomlex/index.html>
- [Schierholz 2001] Schierholz, S.J. (2001). Präpositionalattribute. Syntaktische und semantische Analysen. *Linguistische Arbeiten 447*- Tübingen.
- [Schmid 1994] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44-49.
- [Schmid 1999] Schmid, H. (1999). Improvements in Part-of-Speech Tagging with an Application to German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann and D. Yarowsky (eds), *Natural Language Processing Using Very Large Corpora*. volume 11 of *Text, Speech and Language Processing*. Kluwer Academic Publishers, Dordrecht, pp. 13-26.
- [Schmid 2000] Schmid, H. (2000). Unsupervised Learning of Period Disambiguation for Tokenisation. Internal Report, IMS, University of Stuttgart.
- [Schmid/Fitschen/Heid 2004] Schmid, H., A. Fitschen and U. Heid (2004). SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of LREC-2004*. Lisbon, Portugal.
- [Schulte im Walde 2002] Schulte im Walde, S. (2002), A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, 1351-1357. Las Palmas de Gran Canaria, Spain.
- [Schulte im Walde 2006] Schulte im Walde, S. (2006). The induction of verb frames and verb classes from corpora. To appear in A. Lüdeling and M. Kytö (eds), *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.
- [Schumacher 2004] Schumacher, H., J.Kubczak, R.Schmidt and Vera der Ruiter (2004). VALBU - Valenzwörterbuch deutscher Verben. Tübingen: Gunter Narr Verlag.
- [Sommerfeldt/Schreiber 1983] Sommerfeldt, K. and H. Schreiber (1983a). Wörterbuch zur Valenz und Distribution deutscher Adjektive. Leipzig: VEB Bibliographisches Institut.
- [Sommerfeldt/Schreiber 1996] Sommerfeldt, K. and H. Schreiber (1996). Wörterbuch der Valenz etymologisch verwandter Wörter: Verben, Adjective, Substantive. Tübingen Niemeyer.
- [Spranger 2004] Spranger, K. (2004) *Beyond Subcategorization Acquisition - Multi-Parameter Extraction from German Text Corpora*. in Geoffrey Williams and Sandra Vessier, editors, *Proceedings of the 11th Euralex International Congress* volume 1 pp. 171-176.
- [Zifonun/Hoffmann/Strecker 1997] Zifonun, G., L.Hoffmann, B.Strecker (1997). *Grammatik der deutschen Sprache*. Band 2. Berlin/New York: de Gruyter.