Kobalt: Extension Corpus and Verb Class and Dependency Annotations

Annotation Guidelines

Anna Shadrova
anna.shadrova@hu-berlin.de

July 2019

Contents

1 General Information

Kobalt (Zinsmeister et al., 2012) is a corpus of essays written by learners and native speakers of German in response to the prompt "Did previous generations do better (had a better life) than today's youth?" ("Ging es früheren Generationen besser als der heutigen Jugend?" in the German original). The original Kobalt corpus contains 20 texts each by German native speakers, and learners of German from Belarus and China. All learners scored in the B2 range of the Common European Frameword of Reference for Languages (Council of Europe, 2017). Native speakers were selected from a homogeneous group of speakers, namely the 12th grade of a high-school from a more affluent part of Berlin, Germany. Speaker texts were written manually and transcribed by professional linguists. Texts were collected outside of a classroom setting without any grading. Writing time was limited to a maximum of 90 minutes with no aids allowed during text production.

Annotations of the original Kobalt corpus include target hypothesis layers ZH0, ZH1, and ZH2, cf. Reznicek et al. (2013) and Reznicek et al. (2010), automatically assigned part-of-speech tags (TreeTagger, Schmid 1995) and automatically assigned dependency tags (MaltParser, Nivre et al. (2006), with Foth's 2006 dependency grammar of German). These guidelines describe corrections of dependency parses, partial corrections of part-of-speech tags, and morphological and syntactic verb classification as represented in Kobalt version 1.5.

An additional 42 texts for Chinese and 69 texts for Belarusian learners of German were collected outside of the initial score limit. Those are available as the Kobalt extension corpus in this repository. Annotations as described in this document were added to the data from the base and the extended corpus in the CoNLL format. The Excel data in this repository contains the base text and the target hypotheses with tracked changes from base text to target hypothesis only for the extension corpus. The base text and target hypotheses with tracked changes of the Kobalt base corpus can be accessed through `http://korpling.de/annis`.

The R-object kobalt_data_with_meta.r contains the whole corpus (base and extension) with metadata and various layers of annotation. Verb category annotations are summarized for different types of constructions in verb_cat_new, etymological information is deleted, some errors in the data are corrected, and certain categories are counted for each text (number of modal, auxiliary, finite verbs, participles, nouns, text length). More information about the corpus, data pre-processing, annotation decisions and examples, and metadata can be found in Shadrova (2020, chap. 3).

The original corpus further contained 11 texts by Swedish learners of German. The annotations described in this document were not added to those texts and those 11 texts are not included in this repository.

2 Part-of-Speech Tags and Lemmatization

Part-of-speech tags were corrected for the verb-noun domain, but not generally for pronouns and modifiers. Lemmatization was corrected. This was mainly necessary for newer vocabulary that was not covered by TreeTagger and for partial homonyms/homographs (for example *fallen/gefallen*, where the participle of the two verbs is homonymous).

3 Verb Annotations

3.1 Auxiliary Verbs

Auxiliaries (tagged *aux*) are verbs used for temporal and passive constructions. In German these are *haben* (in auxiliary use – lexical *haben* was tagged as *simple*), *sein* (in auxiliary use – copula *sein* was tagged as *copula*) and *werden* in temporal constructions.

3.2 Constructional Verbs

The tags *infinitive_cx, reflexive_cx, halten_cx,* and *haben_cx* were used to tag lexical verbs in constructional (and frequently non-compositional) use. This includes infinitive constructions such as *das ist schwierig zu sagen* ('it's hard to say') and reflexive constructions such as *es lässt sich sagen; es zeigt sich, dass ...* ('it can be said', '... becomes apparent', literally 'it lets itself say', 'it shows itself that ...'). This is a tentative categorization that was given only to verbs triggering specialized syntactic environments. For example, in *Er fasst sich an den Kopf* 'He is touching his head', *fassen* would not be tagged as a constructional verb.

halten_cx and *haben_cx* tag the constructions *jemanden für etwas halten*, for instance *Ich halte mein Land für glücklich* ('I think of my country as a happy place', literally 'I hold my country for happy') and *es gut/schlecht haben* ('to be in a good/bad situation/place', literally 'to have it good/bad'). Both are rather common in the corpus due to the prompt.

3.3 Copula Verbs

Copula verbs (marked *copula*) connect a subject with a predicate. In German, this can be done using the verb *sein* 'to be', or *werden* 'to become', for instance in *er wird langsam alt*, 'he is slowly getting old, he is aging'. Auxiliary uses of *werden*, as in *Sie wird glücklich sein* 'she is going to be happy' were marked as *aux*. In this annotation, *bleiben* 'remain, stay (active in a function)' was marked as copula if it was used in a copula sense, for instance *Vieles bleibt in allen Zeiten gleich* 'much remains the same through all times', since it is parallel to *Vieles ist n beiden Ländern gleich* 'much is the same in both countries'.

3.4 gehen_cx and gehen_um

Since Kobalt's prompt contains the construction *geht es ihnen besser*, 'are they doing better', literally 'is it going better to them', participants make abundant use of this

construction. Since *gehen* is also a simplex verb with the meaning of 'to go', but occurs much more frequently than other constructions in these specific texts, a differentiation was necessary and the constructional use is marked as *gehen_cx*. *Gehen* further occurs in the construction *es geht um/darum (dass es den folgenden Generationen besser geht)* 'it is about (making it better for the coming generations)', which is marked as *gehen_um*.

3.5 Modal Verbs

Modal (marked *modal*) verbs are a closed class containing the verbs *dürfen, können, müssen, sollen, wollen* and *mögen* in modal use (*ich möchte sagen, dass...*, but not *sie mochten Bücher*, where it would be tagged as simplex/lexical). The tag was further used for *brauchen*, where it is used in a modal (rather than a modifying) syntactic frame (*er braucht das nicht machen*) and *werden* in a modal construction (*Ich würde mir das niemals erlauben; Meiner Meinung nach würde es logisch sein*).

3.6 Modifying Verbs

Modifying verbs are similar to modal verbs, but are complemented by a full infinitive phrase (*er braucht das nicht **zu** machen*). There are only two modifying verb lexemes in Kobalt, *brauchen* and *scheinen*.

3.7 Particle Verb

Split particle verbs (marked *particle*) are complex verbs consisting of a base and a particle that is split in certain syntactic contexts. For example, the verb *auffüllen* 'fill up' would occur as *Sie **füllt** den Wassertank **auf*** in present tense (Präsens) or as *Sie hat den Wassertank auf**ge**füllt* in perfect tense (Perfekt). A particle can be a preposition, an adverb, or an adjective, but is in all cases an existing word or a free morpheme. Particles, unlike prefixes (see next section), are also stressed.

Annotation was performed following a maximizing principle, i.e. anything that could be a particle in the given environment was considered as one. The particle was attached to the lemma in split cases, where STTS normally tags only the base verb.

In case of the coordination of an adverb that could be a particle and a prepositional phrase, the verb was tagged as simplex to avoid issues with duplex lemmatization (for example: *Er läuft hinterher und die Straße entlang*).

This is full list of particle verbs as tagged in Kobalt:

abbauen, abfinden, abgewöhnen, abgrenzen, abhängen, abholzen, ablaufen, ablenken, abnehmen, abreagieren, abrufen, abschalten, abschlagen, absehen, abwägen, anbetreffen, anbieten, aneignen, anerkennen, anfangen, anfertigen, anfreunden, anführen, angeben, angehen, angreifen, angucken, anhäufen, anhören, anklagen, anklopfen, ankommen, anlegen, anlocken, anmelden, anmerken, annähern, annehmen, anpassen, anpregten, anprobieren, anrufen, anschaffen, anschalten, anschätzen, anschauen, anschließen, ansehen, ansetzen, ansprechen, ansteigen, anstellen, anstrengen, antragen, anweisen, anziehen, aufbauen, aufbringen, auffallen, auffordern, aufgeben, aufhalten, aufklären, aufmachen,

aufnehmen, aufpassen, aufpicken, aufregen, aufrufen, aufschlagen, aufschreiben, aufspringen, aufstehen, aufsteigen, aufstellen, auftauchen, aufteilen, auftreffen, auftreten, aufwachen, aufwachsen, aufweisen, aufwerfen, aufziehen, ausarten, ausbeuten, ausbilden, ausbrechen, ausdehnen, ausdenken, ausdrücken, auseinandersetzen, ausgeben, ausgehen, aushängen, ausheilen, aushungern, auskennen, auskommen, auslächeln, ausleben, auslegen, auslösen, ausmachen, ausnutzen, ausprägen, ausreichen, ausrichten, ausrotten, ausrufen, ausruhen, ausschließen, aussehen, aussetzen, aussiedeln, aussprechen, ausstatten, ausstrahlen, aussuchen, austauschen, austragen, ausüben, auswählen, auswerten, auswirken, auszeichnen, ausziehen, beilegen, beitragen, bekanntmachen, darstellen, durchbrechen, durchführen, durchsehen, durchsetzen, einbeziehen, einbringen, einfallen, einführen, einfüllen, eingehen, eingrenzen, einkaufen, einladen, einleben, einnehmen, einprägen, einrichten, einschalten, einschätzen, einschlafen, einschließen, einschränken, einschreiben, einsehen, einsteigen, einstoßen, einteilen, eintreten, einwirken, fernsehen, fertigwerden, festhalten, festlegen, feststehen, feststellen, fortsetzen, freilassen, herangehen, herausfinden, herstellen, herumtollen, herumweinen, herunterladen, hervorbringen, hervorkommen, hervorrufen, hervortreten, hineinversetzen, hinfügen, hinreißen, hinweisen, hinzufügen, kennenlernen, loszuwerden, mitbringen, miteinbeziehen, miterleben, mithelfen, mitkommen, mitmachen, mitnehmen, mitreißen, nachdenken, nachfolgen, nachkommen, nachschlagen, nachvollziehen, näherkommen, rausfinden, rausflüstern, rausgucken, rausschmeißen, rumtanzen, runterladen, schwerfallen, standhalten, stattfinden, stehenbleiben, stillstehen, teilnehmen, übereinstimmen, umbringen, umgehen, umsetzen, umziehen, unterordnen, verlorengehen, voranbewegen, voraussagen, vorbeilaufen, vorbereiten, vorfinden, vorgehen, vorherrschen, vorkommen, vorleben, vorliegen, vorschlagen, vorschreiben, vorstellen, vortäuschen, vorweisen, vorwerfen, vorziehen, wahrnehmen, wegbringen, wegdenken, weglassen, wegnehmen, wegwerfen, weiterbilden, weiterentwickeln, weitergeben, weitergehen, weiterspielen, widerspiegeln, zugeben, zugehen, zugestehen, zugutekommen, zuhören, zunehmen, zurechtfinden, zurechtkommen, zurückdenken, zurückführen, zurückgehen, zurückgreifen, zurückkehren, zurückkommen, zurückliegen, zurückschrauben, zusammenarbeiten, zusammenfallen, zusammenfassen, zusammenfinden, zusammenhängen, zusammenleben, zusammensammeln, zusammensetzen, zusprechen, zusteigen, zusteuern, zustimmen, zutreffen.

3.8 Prefix Verbs

Prefix verbs, unlike particle verbs, are made from a bound or free morpheme and a base verb. The prefix is non-detachable and unstressed. Verbs modified with a free morpheme that were used in an ambiguous position (meaning they could either be read as prefix or as particle verbs) were annotated following the canonical German usage. For example in *Die Eltern wollten sie überzeugen*, which could potentially be used as a particle verb (*Sie zeugte über*), but is not canonical in German.

The full list of prefix verbs as they are tagged in Kobalt is as follows:

beachten, beantworten, bedanken, bedauern, bedenken, bedeuten, bedrohen, bedürfen, beeindrucken, beeinflussen, beenden, befassen, befehlen, befinden, befolgen, befragen, befreien, befreunden, befriedigen, befürworten, begegnen, begehen, beginnen, be-

gleitend, begraben, begründen, behalten, behandeln, behaupten, beheben, beherrschen, bejahen, bekämpfen, bekennen, beklagen, bekommen, belasten, belegen, beleidigen, belügen, bemerken, bemühen, benehmen, beneiden, benennen, benötigen, benutzen, benützen, beobachten, bereichern, bereisen, bereiten, berichten, berücksichtigen, beruhen, berühren, beschädigen, beschäftigen, beschleunigen, beschönigen, beschränken, beschreiben, beschuldigen, beschützen, beschweren, beseitigen, besetzen, besichtigen, besiedeln, besinnen, besitzen, besorgen, besprechen, bestätigen, bestehen, bestellen, bestimmen, bestrafen, bestreiten, besuchen, betonen, betrachten, betreffen, betreten, betreuen, betrügen, beurteilen, bevorzugen, bewahren, bewältigen, bewegen, beweisen, bewerben, bewerten, bewirken, bewundern, bezahlen, bezeichnen, beziehen, bezweifeln, einscheiden*, empfinden, entdecken, entfalten, entführen, entkommen, entlassen, entnehmen, entscheiden, entschließen, entschuldigen, entspannen, entsprechen, entstehen, enttäuschen, entwickeln, erahnen, erfahren, erfinden, erfolgen, erfordern, erfreuen, erfüllen, ergeben, ergehen, ergreifen, erhalten, erhöhen, erholen, erinnern, erkämpfen, erkennen, erklären, erlangen, erlauben, erläutern, erleben, erledigen, erleichen, erleichten, erleichtern, erleiden, erlernen, erleuchten, ermahnen, ermöglichen, ermorden, ermutigen, ernähren, erneuern, eröffnen, erreichen, erscheinen, erschweren, ersetzen, ersparen, erstaunen, erteilen, ertragen, erwachsen, erwähnen, erwarten, erweitern, erwerben, erzählen, erzeugen, erziehen, erzielen, fallen|gefallen, gebrauchen, gefallen, gehören, gelangen, gelangen|gelingen, gelingen, genießen, geraten, geraten|raten, geschehen, gestalten, gestehen, gewähren|währen, gewährleisten, gewinnen, gewöhnen, hinterlassen, kennzeichnen, missbrauchen, missfallen, schlussfolgern, überfordern, übergeben, überholen, überlasten, überleben, überlegen, übernehmen, überqueren, überraschen, überschreiten, überschütten, übersehen, übersetzen, überstehen, überwiegen, überwinden, überzeugen, überzeugt, umfassen, umgeben, unterbrechen, unterdrücken, unterhalten, unterliegen, unternehmen, unterrichten, unterschätzen, unterscheiden, unterstreichen, unterstützen, verabreden, verabschieden, verachten, verallgemeinern, veraltern, verändern, veranlassen, veranschaulichen, veranstalten, verbergen, verbessern, verbieten, verbinden, verbrauchen, verbreiten, verbringen, verdanken, verderben, verdeutlichen, verdienen, vereinfachen, vereinheitlichen, vereinigen, vererben, verfassen, verfeinern, verfolgen, verfügen, vergeben, vergehen, vergessen, vergeuden, vergewaltigen, vergleichen, vergnügen, vergrößern, verhaften, verhalten, verhätscheln, verheimlichen, verheiraten, verhindern, verhungern, verkaufen, verkehren, verkleinern, verlangen, verlängern, verlangsamen, verlassen, verleben, verlegen, verleichten, verleihen, verleiten, verletzen, verlieren, vermeiden, vermissen, vernachlässigen, verneigen, vernetzen, vernichten, vernünftigen, veröffentlichen, verpassen, verpesten, verpflichten, verringern, versammeln, verschaffen, verschärfen, verschlechtern, verschließen, verschmutzen, verschulden, verschwenden, verschwinden, versehen, versetzen, versorgen, versprechen, verspritzen, verspüren, verstärken, verstecken, verstehen, verstoßen, versuchen, vertauschen, vertiefen, vertrauen, vertreten, verunreinigen, verursachen, verurteilen, vervollkommnen, verwahren, verwandeln, verwechseln, verweigern, verwenden, verwerten, verwickeln, verwirklichen, verwöhnen, verwundern, verzichten, verzweifeln, vollziehen, vorbehalten, widerlegen, widersetzen, widerspielen, widersprechen, wiederholen, zerbrechen, zerreißen, zerstören

**einscheiden* was understood as a learner variant of *entscheiden* and thus categorized as a prefix verb.

3.9 Simplex Verbs

All verbs that are not in any of the other categories are marked as simplex verbs (tagged *simple*).

3.10 Support Verbs

This tag should be viewed with caution and validated if it were to be used as a crucial element of an analysis.

Support verbs (tagged *support*) are lexical verbs used to form support verb constructions with an NP, AdvP or AdjP complement and (typically) a non-compositional meaning, such as *eine Rolle spielen, ins Auge springen, verrückt spielen.* Support verbs were annotated by these two factors without further validation through a dictionary or inter-annotator agreement scores. Counting support verbs as lexical/simplex verbs instead should be safe.

3.11 Passive

All verbs in passive constructions (including the auxiliary) were additionally tagged as passive. Thus, to count the number of passive constructions, the number of passive auxiliaries should be counted rather than the number of passive tags. The passive tag was added to the verb category separated by an underscore, e.g. *aux_pass* and *simple_pass.* This is not an ideal representation and should be separated into its own tier in future versions.

3.12 Etymological Origins

This is an incomplete and legacy annotation and should not be considered for crucial parts of an analysis.

Parts of the corpus contain information on etymological origins of verbs, tagged *ang, fra,* and *lat* for English, French, and Latin origin. This was added to the verb category tags separated by an underscore (e.g. *simple_ang* or *reflexive_cx_ang*).

4 Dependency Parses

Dependency parses were manually corrected with the following changes to Foth's 2006 dependency grammar of German:

4.1 Subjects

Subjects were assigned as dependents of the lexical verb of a construction, not the auxiliary or modal verb. The dependency chain would be auxiliary → lexical verb →

{all objects, subjects, and adverbials}. This is unlike Foth's grammar, in which only the subject is assigned dependency of the auxiliary due to congruence in number and person. This can be reverse-engineered if necessary.

4.2 Coordination

Coordinated dependencies were all marked by function rather than coordination labels. Thus a verb may take several accusative objects. In Foth's grammar this would be marked as OBJA → {a number of coordinating conjunctions and/or listed arguments all marked as KON} → the final argument marked as CJ. The labels KON and CJ would be given identically for any type of dependency, for example accusative objects, adverbs, or clause-type complements. In this annotation, they are marked according to their function, for example OBJA, ADV, or OBJC.

4.3 PP-Attachment

Prepositional phrases (PP) were assigned to the verb (rather than an NP) whenever they could plausibly be put in the German Vorfeld (position before the finite verb in a German sentence). Choices have not been validated through inter-annotator agreement scores.

4.4 PP or OBJP

Prepositional phrases were marked as prepositional objects (OBJP) whenever they were either obligatory or would canonically be considered governed by the verb. This was not validated through inter-annotator agreement scores. OBJP was also assigned to complex predicates formed through a prepositional phrase, as in the example *Einige davon sind eingebildet und sozial ohne Pflichtbewusstsein* ('Some of them are arrogant and show no sense of social responsibility'). In this case, *eingebildet* is a predicate, as is *sozial ohne Pflichtbewusstsein*. However, Foth's grammar does not allow for prepositional predicates, prepositions are either assigned as head of PP or OBJP. However, the copula *sein* does not take objects, it only assigns predicates. In order to still be able to tell this kind of prepositional phrase from the modifying PP in*Sie machen Urlaub in Bayern* ('They spend their vacation in Bavaria'), OBJP was assigned as a label in these cases, too.

4.5 Participles

Participles can occur as verb argument structures or deverbal adjectival predicates. For example, in the sentence *Die Frauen oder auch Mädchen wurden als den Männern und Jungs untergeordnet betrachtet* ('Women and girls were seen as inferior/subordinate to men and boys'), *untergeordnet* can be analyzed as a deverbal adjective with an inherited argument structure, or as a verbal modifier (similar to *Die Sache wurde als beschlossen betrachtet* ('The issue was seen/viewed as decided'). In this annotation, participles carrying a deverbal argument structure were annotated as state passives (dependent on an auxiliary if there is one), while participles not showing their argument structure were

treated as predicates (dependent on a copula if there is one) or adjectival modifiers. For example, in *Sie sind den Männern und Jungs untergeordnet*, *sind* would be tagged as an auxiliary in the verb category and *untergeordnet* would be tagged as an AUX in the dependency (AUX marks verbs dependent on auxiliaries). In *Das ist eine untergeordnete Frage*, *untergeordnet* would be tagged as an ADJA dependent on the NP. n *Diese Frage ist untergeordnet*, *ist* would be tagged as a copula in the verb category tier and *untergeordnet* would be tagged as *PRED* in the dependency tier.

References

Council of Europe (2017). Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors. *Provisional Edition*.

Foth, K. A. (2006). Eine umfassende Constraint-Dependenz-Grammatik des Deutschen.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Reznicek, M., Ludeling, A., and Hirschmann, H. (2013). Competing target hypotheses in the Falko corpus. *Automatic treatment and analysis of learner corpus data*, 59:101–123.

Reznicek, M., Walter, M., Schmidt, K., Lüdeling, A., Hirschmann, H., Krummes, C., and Andreas, T. (2010). Das Falko-Handbuch: Korpusaufbau und Annotationen. *Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin*.

Schmid, H. (1995). Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Shadrova, A. (2020). *Measuring coselectional constraint in learner corpora: A graph-based approach*. Univ.-Dissertation, Humboldt-Universität zu Berlin.

Zinsmeister, H., Reznicek, M., Brede, J. R., Rosén, C., and Skiba, D. (2012). Das Wissenschaftliche Netzwerk" Kobalt-DaF". *Zeitschrift für germanistische Linguistik*, 40(3):457–458.