A06-1 General information about Project A06

A06-1.1 Project title: Disentangling cross-linguistic and language-specific aspects of register variation

A06-1.3 Project leaders

Prof. Dr. Elisabeth Verhoeven
Inst. f. deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin
elisabeth.verhoeven@hu-berlin.de
Telefon: +49 30 20 93 97 96
Telefax: +49 30 12 34 34 56

Prof. Dr. Aria Adli
Romanisches Seminar
Universität zu Köln
Albertus-Magnus-Platz, 50923 Köln
aria.adli@uni-koeln.de
Telefon: +49 2 21 4 70 44 48

A06-2 Summary

A06 aims at investigating cross-linguistically manifested aspects of register, which are supposed to be reflexes of universal communicative behaviour, differentiating them from idiosyncratic language-specific register variation. The study will focus on grammatical means of information packaging that are likely to be register-sensitive. In particular, we will investigate word order variation and variation in referential expressions as dependent on properties of the communicative situation. For instance, certain non-canonical word orders (depending on the potential of a particular grammar) seem to be especially employed in informal registers, resulting from a greater adaptation to interactional needs. Hence, information structural devices reducing syntactic compactness, such as right-dislocations or certain fronting operations, are expected to occur more frequently (across languages) in registers of spontaneous speech. On the other hand, syntactic variation can also show an arbitrary and conventionalized variant-register relation, which is expected to be language-specific, e.g. in French wh-in-situ is restricted to informal language (Adli, 2015).

Cross-linguistic questions regarding general mechanisms of register variation have been addressed only rarely (e.g. in Biber, 1995). Especially smaller and less researched languages do not figure prominently in register research. A06 will address both research gaps and investigate and compare register variation in three typologically diverse languages with widely different socio-cultural situations and research histories, namely German, Persian, and Yucatec Maya. A leading question of A06 will be whether languages with large distance between registers, possibly on a par with diglossia (e.g., Persian) are more prone to develop idiosyncratic form-register associations than languages with less pronounced registers (e.g., Yucatec Maya). In addition to the investigation of existing corpora we will build parallel materials in the three languages collected by the same methods, namely spontaneous speech recordings in different communicative settings, judgments on the appropriateness of syntactic variants in specific situative contexts, as well as a situative classification task (with certain parallels to the matched guise technique). For these controlled parallel studies, we will focus on two register parameters, which have been proven relevant in register studies in many languages, namely distance between the interlocutors (linked to formality) and modality (written vs. spoken). We will measure to what extent (a) the conditional probability of particular syntactic variants depends on register and (b) speakers are aware of the association of certain variants with particular types of situative contexts. In order to analyze intra-speaker variation and compare production and perception data, all participants will take part in the full design, i.e. recorded in the different settings and participate in judgment studies, both in auditory and in written modality.

The findings of A06 will contribute to the central question of the CRC concerning the role of register variation in grammar (Area A). In particular, this project will add findings from different language situations (from diglossia in Persian to incipient written-spoken variation in Yucatec Maya) to the discussion. The types of phenomena we are dealing with involve differences in the frequency of constructions that are available across registers. The crucial question in syntactic instances of variation is to identify the layer of variation: is the observed variation between registers due to differences at the level of expression (e.g. likelihood of dropping uniquely identifiable subjects), or due to differences at the functional level (e.g. likelihood of uniquely identifiable subjects)? Given these two layers of variation, the interplay between register and information structural context in determining the occurrence of syntactic variants will play a central role in A06.

A06-3 Research Rationale

A06-3.1 Current state of research and preliminary work

The roots of syntactic variation between registers are, at least partly, functional in nature: speakers optimize their linguistic behaviour in order to achieve particular situation-specific goals (Biber, 1995; cf. also S. C. Levinson, 2006; Trudgill, 2011). This is reflected in properties of register that are **cross-linguistically consistent** although they independently emerged in different speech communities. For instance, it comes as no surprise that structural complexity may correlate with the amount of attention paid to speech (Givón, 2009). Crucially, such aspects of linguistic behaviour are independent of a particular language and are related to general principles of communication (see the "Effort Code" in Gussenhoven, 2004). On the other end of the spectrum, we observe highly **language-specific mappings** between syntactic structures and registers that cannot be straightforwardly explained in functional terms. For instance, in Turkish, passive voice is mostly confined to broadcasts and official speeches (Göksel and Kerslake, 2005), or in French wh-in-situ is restricted to informal language (Adli, 2015). In such cases, the specific association of particular variants with certain registers is not cross-linguistically predictable.

The main aim of A06 is to investigate these two classes of phenomena with particular emphasis on their contribution to the construction of registers. We will focus on syntactic phenomena that are known to be register-sensitive in different languages in order to disentangle language-specific and cross-linguistic components of register variation. The comparison between languages allows us to tackle the question whether some aspects of syntactic variation are cross-linguistically associated to certain registers. For instance, the preference to exploit word order options (depending on the potential of a particular grammar) in informal registers may be functionally motivated to some extent, resulting from a greater adaptation to interactional needs. Hence, information structural devices reducing syntactic compactness, such as right-dislocations or certain fronting operations are expected to occur more frequently, across languages, in registers of spontaneous speech (see also A03). On the contrary, syntactic variation can also show an

arbitrary and conventionalized variant-register relation, i.e. an instance of form-to-function association that does not essentially differ from the association of lexical forms with particular registers, and is connected with an indexical field (Eckert, 2008). This part of the variation is expected to be language-specific.

Cross-linguistic work on register variation, especially regarding smaller languages, is still in its infancy. Biber's (1995) work on register variation in four widely divergent languages (English, Korean, Somali, Tuvaluan) is ground-breaking in this respect. Using a bottom-up approach, this study shows striking similarities across these languages concerning certain register dimensions. Two dimensions are particularly robust across languages, namely clausal/oral vs. phrasal/literate discourse and narrative vs. non-narrative discourse, to use terminology from Biber (2014). Next to this cross-linguistic approach, there is a growing body of contrastive work, which is highly relevant for our endeavour (e.g. Dimroth, Andorno, et al., 2010; Auer and Maschler, 2013; S. Neumann, 2013, 2014; Kunz and Lapshinova-Koltunski, 2015). Part of this work is emerging from the interest of understanding translation processes and is usually corpus-based, treating well-studied languages such as German and English. Another line of comparative and contrastive register studies has evolved in language acquisition (L1 and L2) research, again with a focus on well-studied, mostly Germanic and Romance languages (cf. Hickmann and Hendriks, 1999; M. Carroll and M. Lambert, 2003; Stutterheim and M. Carroll, 2005, among others). These studies uncover fine-grained differences and commonalities between these languages in the register-specific use of e.g. cohesion devices, anaphoric linking devices or subject choice yielding a fine-grained picture of micro-variation with respect to closely related languages. Turning to less or under-researched languages, register studies are still rare, and these languages have not been systematically integrated in cross-linguistic register research (with the notable exception of Biber (1995)). Hence we largely lack a more profound knowledge of cross-linguistic properties of registers variation as well as systematic studies on the interaction of typological features with certain registers.

In order to be able to focus on cross-linguistic properties of register variation, we will base our investigation on (a) well-researched parameters of register distinctions, namely **formality** and **modality**, and (b) instances of syntactic variation that are already known to be register sensitive in many languages. Here we will focus on syntactic means of coding **information structure**, focusing on word order variation and referential expressions. We believe that these phenomena are well-suited to investigate register variation. Previous research in several languages has shown that many instances of information packaging favor or disfavor certain constructions, rather than obligatorily requiring specific syntactic structures. For example, topic shift in Spanish subject pronouns increases the rate of overt pronouns but it does not require an overt realization (Adli, Forthcoming). Likewise, narrow focus in French increases the probability of clefting, but it does also occur in canonical word order (Dufter, 2008).

Register studies have focused on diverse parameters. For instance, the difference between spoken and written modality with regard to complexity is a recurrent issue in register research (e.g. Halliday, 1979; Miller and Weinert, 1998; Maas, 2010, among many others). More recent studies have depicted a more diversified picture. Biber (2014) observes that the distinction between oral vs. literal production is cross-linguistically manifested through similar linguistic structures. Speech is characterized by a preponderance of "clausal" discourse, the use of pronouns, verbs and adverbs, whereas writing heavily relies on "phrasal" structures, including the use of recursive nominal embedding (e.g., S. Neumann, 2013, 2014). Another famous register parameter that has been studied and found relevant in many languages is the contrast between formal and informal interaction. For instance, Paolillo (2000) distinguishes between different spoken registers in Sinhala that vary in formality, the formal variety showing a higher complexity in terms of the coding of grammatical features than the less formal variety.

An important issue in cross-linguistic generalizations about registers is the varying degree of **register diversity across languages**. We already know about the correlation between register use and certain external (social) variables, considered to be a general principle in the sociolinguistic literature. Famously, women use the standard variant more often than men for change in progress above the level of awareness and more often the incoming, vernacular variant for change below the level of awareness (Labov, 2001). However, we have much less cross-linguistic knowledge about register use in relation to internal variables, e.g. whether certain linguistic variables are more prone to register variation across languages. On this background, the role of register diversity is interesting to investigate. For instance, are languages with large distance between registers more prone to developing idiosyncratic form-register associations, possibly on a par with the development of diglossia? Further aspects that have been shown to play a role in the establishment of genres and registers include the existence of (official) norms, which come along with the use of a language

in education, media, etc, and also the prestige associated to the vernacular of a certain group of speakers, see Modarresi (1978) on the prestige of the vernacular spoken by the inhabitants of the capital Tehran, used by speakers from the province in their *formal* speech.

Given the overall quantitative and probabilistic conception of register as advocated in the CRC (cf. Section 1.2.1), a cross-linguistic study including a larger typologically and genetically balanced sample of languages seems not feasible at present. A06 proposes **a parallel in-depth study** of three languages, namely Ge[rman], Pe[rsian], and Y[ucatec] M[aya], which possess typologically different but comparable morphological and syntactic devices in the investigated domains of information structure (see Section A06-4.2 for details). A06 will examine instances of syntactic variation related to information structure that are already established as register sensitive in these languages. A preference to exploit **word order variation** (depending on the potential of a particular grammar) has been claimed to exist in registers of informal spontaneous speech. In Pe, formal registers are clearly V-final, while postverbal material is frequent in colloquial registers (Frommer, 1981). In YM, canonical VOS sentences are frequently used in written fairy tales, while in oral speech, subjects are almost always left-dislocated (Skopeteas and Verhoeven, 2009). This situation is reminiscent of the use of postfield material in Ge: while in the written language, the use of constituents following the non-finite verb are avoided, spoken registers are much more flexible in this respect (Hartmann, 2013; Féry, 2015). On the other hand, also elaborated registers of written language may show considerable word order variation. For instance, in a corpus study on Ge written language Verhoeven (2015) found that in around 12 percent of passive clauses with transitive verbs, the non-subject precedes the subject. Given this picture, the cross-linguistic expectation is that registers of informal spontaneous speech are more prone to showing word order variation reducing syntactic compactness, such as dislocations beyond the clausal boundaries (e.g. Givón, 1979; Pawley and Syder, 1983; Rühlemann, 2006, among many others), while more elaborated and planned registers show word order variations inside the core clause/CP domain (scrambling into topic or focus positions). The investigation of word order variation is closely related to the issue of syntactic complexity, which has also been a topic in register research as reported above regarding the influence of modality. Verhoeven and Lehmann (2018) explored complexity in terms of embedding at different syntactic projections and found a significant effect of the distinction 'private vs. public speech' on the depth of embedding at the CP, VP, and DP levels for Ge.

Furthermore, we will examine the usage of **referential expressions** as register-dependent. Variation in the choice of referential expressions has a long research tradition in variationist sociolinguistics, which is corroborated by the high-frequency nature of this phenomenon, especially with regard to the grammatical subject. One case in point is pronominal drop, which is strongly influenced by reference continuity as shown in many Spanish dialects (e.g. R. Cameron, 1992), but also in other languages, among them Pe (Haeri, 1989). Not surprisingly, the same holds for topic continuity (Adli, 2011). However, we assume that the discourse patterns of topic continuity correlate with register and the type of discourse. Narrative texts are likely to have longer stretches with one topic, while spontaneous dialogues come with fast turn changes and shifting topic references of local persons (i.e. the discourse participants). At the same time, local person references follow different rules with regard to topic continuity and pronominal drop than third-person references (Adli, Forthcoming). Most variationist studies have explored the variation between null and overt subject pronouns with Spanish varieties spoken in the US by native or heritage speakers, and without taking register into account. In addition, we are lacking insights from language comparison.

In sum, the syntactic phenomena described are promising domains for a cross-linguistic study of register variation. The PIs of A06 have ample experience in the study of the object languages, especially with regard to the crucial information structural topics. Furthermore, both PIs are very experienced in conducting cross-linguistic experimental research as the one proposed in this project (e.g. Verhoeven, 2010; Adli, 2011; Verhoeven, 2014; Temme and Verhoeven, 2016). The collaboration between the HU Berlin and the University of Cologne in this subproject creates a fruitful synergy on various levels: First, it combines the expertise of PI Verhoeven in typology and syntax with the expertise of PI Adli in variationist approaches to syntax. Second, it connects the infrastructure of the sociolinguistic lab at the University of Cologne with the ZLab at HU Berlin for the empirical aspects of this research. Third, the PIs' research background in different languages (Pe, Ge, YM) are combined for the comparative orientation of the project. Finally, PI Adli has been involved in the coordination of the previous proposal for a research group while he was at the HU Berlin and has continued to be involved in the register project after changing to the University of Cologne in 2014.

A06-3.2 Project-related publications by participating researchers

Peer-reviewed articles and books

Adli, A. (Forthcoming). Topic chains in dialogues. *Journal of Pragmatics* (Prominence in Pragmatics).
Skopeteas, S. & E. Verhoeven (2009). The interaction between topicalization and structural constraints: Evidence from Yucatec Maya. *The Linguistic Review* 26.2–3, 239–259.
Adli, A. (2011). Gradient acceptability and frequency effects in information structure: A quantitative study on Spanish, Catalan, and Persian. Habilitation. Universität Freiburg.
Verhoeven, E. (2014). Thematic prominence and animacy asymmetries: Evidence from a crosslinguistic production study. *Lingua* 143, 129–161.
Adli, A. (2015). What you like is not what you do: Acceptability and frequency in syntactic variation. In: *Variation in Language: Usage-based Vs. System-based Approaches*. Ed. by A. Adli, M. G. García & G. Kaufmann. de Gruyter Mouton, 173–199.
Verhoeven, E. (2015). Thematic asymmetries do matter! A corpus study of German word order. *Journal of Germanic Linguistics* 27.1, 45–104.
Temme, A. & E. Verhoeven (2016). Verb class, case, and order: A cross-linguistic experiment on non-nominative experiencers. *Linguistics* 54.4, 769–814.
Adli, A. (2017). Variation in style: Register and lifestyle in Parisian French. In: *Selected Papers from the 8th International Conference on Language Variation in Europe (ICLaVE 8)*. Ed. by I. Buchstaller & B. Siebenhaar. John Benjamins, 157–171.
Verhoeven, E. & N. Lehmann (2018). Self-embedding and complexity in oral registers. *Glossa: A Journal of General Linguistics* 3.1, 93.

A06-4 Project plan

A06-4.1 Objectives

A06 will be situated within Area A of the proposed CRC addressing the overall question of how register knowledge relates to grammatical aspects of linguistic knowledge (QA). It does so by specifically taking a comparative viewpoint and addressing aspects of the universal and language-specific nature of register variation. We will test our hypotheses by directly comparing three typologically different languages with widely different register distinctions through the parallel application of the same methods. The main aim of A06 is to investigate two classes of phenomena in their contribution to the construction of registers, namely functionally based vs. conventionalized associations between structural variants and registers. Research goals 1–4 jointly address the specific research questions of Area A of the CRC concerning the nature (QAi) and the choice (QAii) of specific syntactic alternants with respect to register, and the consequences for the integration of register in a model of grammar (QAiii).

Research goal 1: *Cross-linguistic vs. language-specific properties of register*

> Which aspects of syntactic variation are cross-linguistically associated with register and which aspects are language-specific?

Languages greatly vary in register diversity. Some languages (Pe) show more salient differences between registers than others (Ge), in spite of both showing similar properties regarding language use in education, administration, media etc. Still other languages (YM) are mainly used for oral communication with only incipient literal use (see QBi).

Research goal 2: *Impact of register diversity*

> What is the impact of differences in register diversity and normative aspects on cross-linguistic similarities and differences in register variation? Are languages with palpable distance between registers more prone to developing idiosyncratic form-register associations, possibly on a par with the development of diglossia?

In order to disentangle language-specific and cross-linguistic components of register variation we will focus on syntactic phenomena related to the encoding of information structure. The phenomena to be considered include (a) word order operations reducing syntactic compactness, such as right- and left-dislocations and (b) the choice of referential expressions as pronominal or null.

Research goal 3: *Register dependence of information-structural devices*

> Is there a cross-linguistic association of structural devices reducing syntactic compactness with informal spontaneous speech in contrast to formal speech and written language, where we expect existing word order variation to be more clearly restricted to positions within clausal boundaries? How does variability in the use of referential expressions differ by register within and across languages?

Cross-linguistic studies on register variation are still rare. In accordance with the methodological repertoire of the CRC and in cooperation with INF, we will develop methods for the parallel investigation of register variation across languages involving both language production and perception (cf. QMi):

Research goal 4: *Methods of studying registers across languages*

> Basic to the cross-linguistic study of register variation is the parallel application of the same methods in order to maximally control between-language variation. We will build a Lang*Reg corpus based on guided naturalistic (spontaneous) language production. We will complement production data with perception data collected through a gradient judgement study and a situative classification task on the association of syntactic variants with specific situative contexts.

A06-4.2 Sample languages and cross-linguistic categories

We will investigate three languages with widely different socio-cultural situations regarding the occurrence of registers and the distance between them. In particular, the languages differ on parameters such as orality/literacy, community size and social diversification: (a) For many Pe varieties in Iran, formal and colloquial registers represent a situation of diglossia with differences at all linguistic levels (Ferguson, 1959; Modarresi, 1978). The language has a high degree of literacy and a rich written tradition (elaborate norms for written vs. spoken varieties; use in administration, education, and media); corpora of written language are available; the only syntactically annotated conversational corpus of Tehrani Pe (as part of sgs) has been created by PI Adli (Adli, 2016); (b) Ge is similar regarding the richness and diversification of the language; however registers seem to be less distinctive compared to Pe; large corpora of both spoken and written language are available; (c) YM (Mayan, Mexico, 700.000 speakers according to 2012 census) is a language with less pronounced register contrasts, mostly used in oral communication in rural communities with low literacy in the population (Pfeiler and Zámišová, 2006); YM has recently been introduced to the education system, it has an evolving literal use (and norms in the process of being established by academies only in recent years). A corpus of spoken and written language, created by PI Verhoeven and colleagues is available.

The sample languages provide comparable categories in the grammatical domains to be investigated. We present them shortly, with emphasis on the relevant aspects for this project. YM is a strictly head-marking and head-initial language with basic VOS order. It has an articulated left periphery with topic and focus positions (Skopeteas and Verhoeven, 2009; Verhoeven and Skopeteas, 2015). Ge and Pe are SOV-languages, whereby Ge, famously, has a V2 property in main clauses, creating a pre-field that hosts by default the subject constituent or a constituent bearing a salient discourse feature (topic or focus). In all three languages, it is possible to find clause-external material either at the left side of the clause (e.g., hanging topics) or following the clause (e.g., afterthoughts). Ge and Pe are scrambling languages, allowing to rearrange the clausal constituents depending on the information structural domains (see Frey 2006; Haider 2010 on Ge; Karimi 2005 on Pe). Two classes of word order phenomena must be distinguished, which may show different types of sensitivity to register. Descriptively, we introduce the dimension of compactness for this purpose: syntactic operations that are used for the creation of information structural domains inside the clause are part of the compact strategy. Syntactic entities such as hanging topics and afterthoughts, which are not integrated in the clause structure are non-compact.

(1) Compactness

> **compact strategies:** scrambling, focus fronting, etc.

non-compact strategies: hanging topics, afterthoughts, dislocations, etc.

As for the expressions of (pronominal) reference, YM and Pe are both *pro-drop* languages with frequent null arguments in both subject and object functions. In YM, person is marked in the form of cross-reference markers on the verb (for subjects and objects); the additional occurrence of personal pronouns signals emphasis. In Pe, a language with subject-verb agreement, but no object agreement, both arguments can be dropped, however under different conditions (Sato and Karimi, 2016). Object drop is restricted to colloquial speech. In Ge, topical 3rd person arguments can be dropped when occurring in the sentence-initial position of a declarative clause (Trutkowski, 2016). In the first funding phase, we will concentrate on the behaviour of null subjects, potentially extending the investigation to object drop.

(2) Null subjects

Yucatec Maya/Persian: yes

German: restricted (only by topic drop)

A06-4.3 Work Packages

A06 is divided into a series of work packages which focus on different work steps in order to achieve Research goals 1–4. WP1–2 address the register parameters formality and modality in production; WP3 deals with register perception. WP4 contains the creation and annotation of a cross-linguistic register corpus, necessary for WP1–3. WP5 evaluates the language-specific results of WP1-3 from a cross-linguistic perspective.

WP1. Formality: registers of oral communication

Background. Registers of oral communication vary along a multitude of parameters, which have been shown for individual languages to shape their properties (Biber, 1995). Formality is among the most researched parameters in this respect influencing speech on several linguistic layers. Formality distinctions are present in our three sample languages, however to different degrees. The degree of formality correlates with several aspects such as the social characteristics of speaker and hearer (age, social position, etc.), their relationship (acquaintance, differences in age, social position, gender, etc.), as well as culture-specific aspects (cf. the prominent distinction between spaces considered to be public vs. private in Iran).

Method. We will investigate registers of spoken language that differ along the parameters associated with formality. Participants will be recorded in two different dialogue situations, namely between well-acquainted and between unknown interlocutors. Each participant will be recorded in both settings, in order to analyze intra-speaker variation (for details see WP4a).

WP1a – Word order indices. We will calculate quantitative indices based on which we can globally assess the variation between registers and between speakers, see (3). The distinction introduced in (1) is crucial for estimating the role of register in different classes of word order phenomena. We expect that more formal interactions will employ devices of the compact type and will avoid devices of the non-compact type.

(3) Word order indices

a. **canonicity index**: the conditional probability of sentences with deviations from canonical word order out of the total sentences in which these deviations are possible.

b. **compactness index**: the conditional probability of sentences with instances of non-compact configurations (hanging topics, afterthoughts) out of the total sentences in the corpus.

WP1b – Word order variation. A06 will conduct variationist analyses of selected syntactic structures (compact and non-compact devices) that appear in particular information structural configurations. For instance, what is the likelihood of right dislocation, as compared to placement of the respective argument within the sentence core? We know that right-dislocation is a communicative means of signaling background information (which does not mean that background information cannot appear within the sentence core). The degree to which speakers tend to make use of "visible" syntactic means to signal such packaging to the hearer might well be register-dependent. For the purpose of efficiency, we will restrict annotations to the categories of interest and only for the relevant subset of tokens. Our annotations will follow standards

developed within previous related projects: for information structure Götze et al. (2007), enriched by the guidelines developed for sgs Adli (2011). Our main aim is to figure out whether the influence of register on the use of particular syntactic constructions is direct, or (partly or fully) mediated by the context (i.e., the contexts that license particular syntactic constructions are more frequent in particular registers). See Figure A06-1 for an illustration of the three factors and their conditional dependencies via directed acyclic graphs. As an example, assume the following situation: postverbal objects occur more frequently in non-formal than in formal registers in Pe; furthermore, word order is influenced by the context: postverbal objects are more likely with given objects than with new objects. The interesting question is whether formality influences the choice of postverbal objects (left graph), or the frequency of contexts with a backgrounded object (middle graph), or both (right graph). In order to answer this question, we need an annotation of the sentences with (preverbal or postverbal) objects and an annotation of givenness in terms of (non-)occurrence in the immediate context – possibly refined with further contextual factors that may increase the accuracy of predicting postverbal orders.

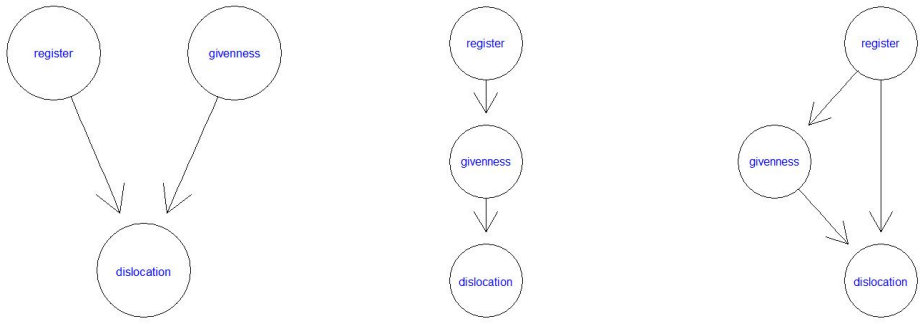


Figure A06-1: Possible models of the interplay of register and givenness regarding dislocation.

WP1c – Referential expressions. Referential expressions reflect how discourse coherence is construed in language. For example, null subjects are favored in contexts of referential continuity, which is even more salient in case of topic continuity. The choice of a subject pronoun or a null subject crucially depends on the properties of the individual grammars; see (2). In this case, register variation is expected to show up with different effects across languages. In a near non-pro-drop language such as Ge, null subjects are expected to appear in informal types of interaction. In a pro-drop language such as YM, subject pronouns are used with emphatic function. These pronouns are particularly frequent in informal spontaneous communication. Our cross-linguistic design allows to assess whether the impact of register varies across languages. Adli (2011) shows that Pe subject pronouns are realized as null in 93% of cases of topic continuity but only in 62% of cases of topic shift in spoken dialogues. Since this variable is generally (at least in Pe and YM) below the level of awareness, we expect parallel changes in pronoun rate from one register to the other across languages. We will annotate POS information for all subjects of finite clauses, and carry out an information-structural analysis for random extracts of each register with regard to this high-frequency variable.

WP2. Modality: spoken vs. written communication

Background. As a second register parameter A06 will test modality. In the spirit of the CRC, which conceives register variation as a multi-factorial research object, it would make sense to cross the factor modality with formality. This is however difficult to realize with regard to YM, since this language does not have established writing practices. Modality will be investigated in narratives being a well-established register in all three languages of investigation, both in spoken and written forms.

Method. Per language, we will collect a story in written and spoken form in parallel along the lines depicted in WP4a. In parallel to WP1, we will measure the word order indices in (3) (**WP2.a**) and annotate and measure the word order variables (**WP2.b**) and subject pronouns/null subjects (**WP2.c**). We expect the modalities to show instances of word order variation of different type (non-compact in spoken, compact in written language, cf. Section A06-3.1). Furthermore, we expect a difference between languages with established writing traditions (Ge and, even more so, Pe) and languages with incipient writing practices (YM), such that the difference between registers is larger in the former than in the latter type of language

situation. As for the referential expressions, Ge topic drop is expected to appear in non-formal dialogues but less so in a spoken narrative. It is not expected in the written form of a narrative. In a language where subject pronouns have mainly an emphatic function, such as YM, we expect pronouns to appear more frequently in oral than in written communication.

WP3. Perception of registers

Background. The relationship between judgment and production is a complex one that is influenced by register as well as by syntactic (sub-)optimality. We know from previous research on French wh-questions that acceptability judgments on constructions that are restricted to colloquial speech receive lower ratings, and constructions mostly used in formal contexts receive a higher rating (Adli, 2015). WP3 addresses speakers' consciousness about register meaning and is related to QAii and QCi. Is the knowledge on situative preferences for certain constructions below or above the level of awareness? In order to estimate whether speakers are aware of the association of certain variants with particular types of situative contexts (i.e. casual vs. careful; communicating with friends vs. strangers), we will collect judgment data from the same participants and compare them with their production data. How large is the (mis)match between what the speakers say about their usage and what they actually do in their use of language? Whenever it comes to norms, it is highly informative to understand both (less conscious) spontaneous speech and (conscious) judgments. Mismatches would be particularly insightful because they can be a window to differences between overt and covert prestige (Trudgill, 1974).

Method. We will combine an acceptability judgment task with a classification task in order to test the structural variants described above, both in spoken and in written modality. We can build on the computer-based gradient acceptability test for written and auditory stimuli already developed (see Adli, 2011). In practice, the informants will be asked to imagine the situative contexts of WP1 and WP2, and make a nuanced acceptability judgment of the target sentence with regard to those situative contexts. Crucially, the target sentences for the spoken contexts will be presented auditorily (on a headphone), while those for the written context will be presented in written form. In order to assess the situative context that the informants imagine, they will also carry out a classification task (with certain similarities with the matched guise technique, W. E. Lambert (1967)). We will ask participants to choose for each construction the most and the least appropriate situative context taken from a predefined set of situations. In order to construct appropriate test material, we will first conduct queries in existing corpora for Pe (sgs), Ge (several), and YM. The outcome of these studies will help us to differentiate functionally motivated and conventionalized constructions. We expect that conventionalized mappings between constructions and registers will show up in the form of robust correlations between form and register and be highly language-specific.

WP4. Lang*Reg corpus

Background. A corpus that contains different registers from the same speakers – including spoken and written language – and which is in addition multilingual, will be a unique data source. We can build on the experience with the sgs corpus, which is multilingual and contains production, judgment, and social data from every speaker. The aim of WP4 (see QMi) is the creation of the Lang*Reg corpus that will be used for the studies of WP1–WP3. It will allow for generalizations estimating the variation within and between speakers (as a random factor).

Data collection. Every participant will participate for 40 min. in four variants of language production (see below), for 30 min. in two variants of a judgment task combined with a classification task (see WP3), and for 25 min. in a social questionnaire. The latter will be based on elements of the social questionnaire used for the sgs corpus with demographic, economic, and socio-cultural indicators, following international standard classifications (Adli, 2017) .

Existing Corpora. The available corpora for YM and Pe contain some of the relevant annotations. The YM corpus of spoken and written language is morphologically annotated and tagged for part of speech. For Pe, the sgs corpus of spoken language is also transcribed and annotated with regard to the grammatical aspects that we will analyze (type of referential expression, word order). The transcriptions in both the YM corpus and sgs are aligned to sound. For Ge, we have several options, among them DeReKo (Institut für Deutsche Sprache, 2018) and DGD (*Datenbank für gesprochenes Deutsch*, dgd.ids-mannheim.de). Most of the DGD corpora include annotations for an orthographic representation, a lemma level and are enriched with part of speech tags according to STTS. The corpora also include metadata with a classification of discourse event, and speaker documentation. The existing corpora will be used in order to estimate the frequency of the different variants of the syntactic variables described above.

WP4a – Corpus design Lang*Reg. Register variation is determined by a multitude of factors, such as the relation between the speakers, the communicative setting and the purpose of the communication. In order to capture the relevant factors, we need a sample of communicative situations that naturally occur in the language communities. The Lang*Reg corpus is designed to (a) be informative for identifying intra-speaker variation between registers of spoken language; (b) be informative for identifying intra-speaker variation in modality (spoken vs. written language); (c) contain diverse communicative situations, in particular with respect to factors that may affect the coding of information structure; (d) be comparable between languages.

We will collect three types of spoken data (dialogue with friend, dialogue with unknown person, narration/monologue) and one type of written data (narration). The spoken dialogue data is necessary for WP1. We will develop a small story for the dialogue settings, which constitute usual incidents in all three cultural settings, such as a moment-of-fear situation reported to a close friend or relative vs. hitherto unknown person, an illness reported to a close friend or relative vs. hitherto unknown person) and ask 12 speakers per language, consisting of 6 pairs of friends, to perform it in the two different settings (close friend/relative vs. unknown interlocutor). In order to realize the two different settings, we will invite four persons at a time to each recording session, i.e. two (non-acquainted) pairs of friends. During the session each pair of friends will tell each other an (authentic) event, then the pairs are rearranged in order to create pairs of persons that do not know each other, and they will do the same task with regard to another (authentic) event. Thus, having 12 speakers, 2 types of relationships between speakers (speakers know/don't know each other), and 10 minutes per dialogue, we obtain a total of $12 \times 2 \times 10 = 240$ minutes of dialogue data for each language. For WP2, we will collect a spoken narrative (monologue) and a written narrative. This amounts to $12 \times 10 = 120$ minutes of spoken data per language. Furthermore, each person will write a narrative during 10 minutes, amounting to $12 \times 10 = 120$ minutes of total writing time. For reasons of cross-linguistic comparison, we will use a small film story for the narrative settings – adapted to each cultural setting – which the speakers, after watching it, are asked to first retell and then to write down (see also Stutterheim and M. Carroll, 2005). The entire recordings will be transcribed.

WP4b – Annotation. The resulting texts will be annotated in close collaboration with INF, following the needs of WP1–WP2. For the annotation, we will build on the TEI data representation for spoken language developed for sgs (Adli et al., 2018) and choose according to the task the annotation tool, such as ELAN, MMAX2, or WebAnno. For conversion into a common format and for querying and analysis, we will use the corpus environment that has been developed at Humboldt University (SaltNPepper, ANNIS, see INF).

WP5. Cross-linguistic properties of register

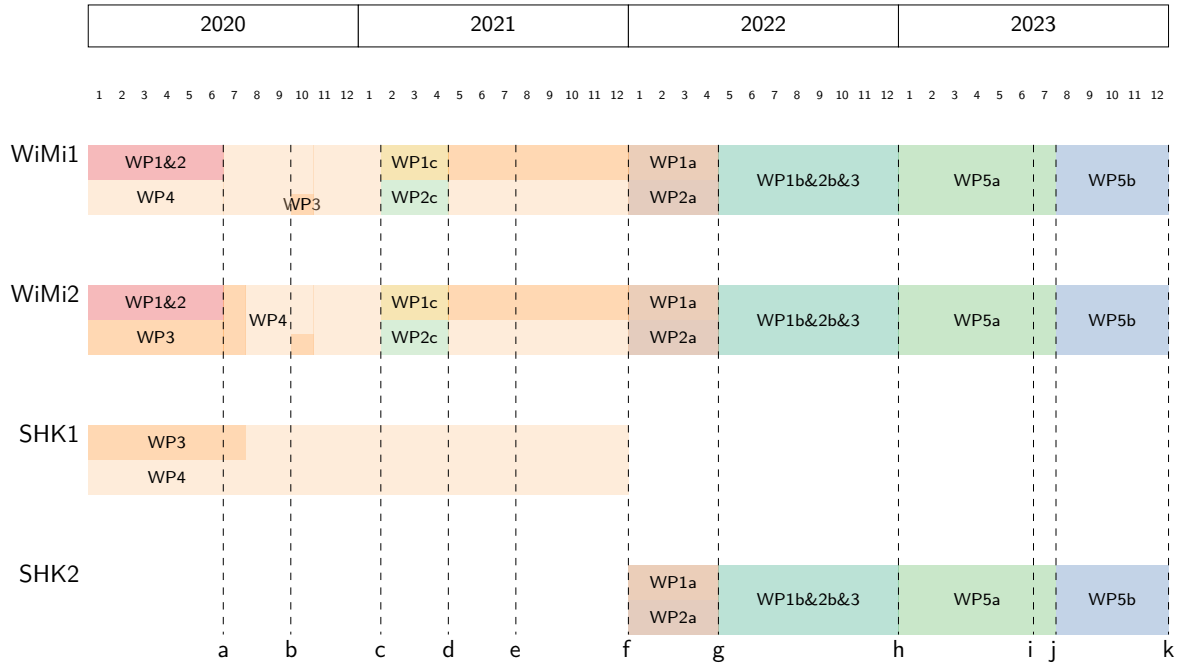
The aim of this work package is to spell out the generalizations of our empirical investigations in a way that is integrated to current theories of the role of register variation in grammar and communication and to combine our findings with insights of related projects in the CRC (QAiii).

WP5a – Cross-linguistic variation and grammatical implementation. In principle, the variants examined in this project, e.g., different linearization options, are structural configurations that are derived within one and the same grammar. Syntactic operations triggered by information structure, e.g., scrambling or focus fronting, are generally considered to be optional, that is there is a residual variation in the occurrence of these phenomena that is not accounted for by the assumption of a discourse feature alone (D. Sankoff, 1988; Adli, 2013). The expected results of our studies will refine the available knowledge regarding the triggers of these linearization options: beyond the discourse features (such as topic or focus), register conditions must be postulated in order to get restrictive descriptions about the necessary and sufficient conditions for the occurrence of the syntactic constructions at issue.

In the viewpoint of the grammatical implementation, the difference between functionally-based and conventionalized constructions is crucial. Conventionalized associations must be learned for a particular language, i.e., they are part of the communicative competence of a speaker. These associations are part of a "register lexicon", i.e., an inventory of associations that is independent of the derivational history of the particular constructions. Functionally-based associations should be explained by general principles of communication, not necessarily represented in grammar.

WP5b – Beyond A06. Results from A02, A03, B02, B04 and C01 are of immediate relevance to our research questions, either since they compare registers cross-linguistically or they address the same or similar phenomena as our project does. A03 deals with word order variation and dislocations in Russian and Czech. We will cooperate with A03 on the evaluation of their results for the cross-linguistic dimension of register variation. C01 investigates the acquisition of registers in multilingual situations and also addresses

dislocation phenomena. We will integrate the results of all projects with a comparative or cross-linguistic focus leading to a more comprehensive picture of cross-linguistic register variation. Furthermore, A06 plans a workshop entitled "Register and word order" in mid 2023, jointly organized with A03 and C01, to specifically discuss results on this topic from the various perspectives present in the CRC (diachronic, typological, acquisition).



a Variable exploration for WP1-3
b Data collection, initial annotation Ge
c Data collection, initial annotation Pe, YM
d Referential expressions in production
e Analysis judgment data
f Cross-linguistic paper on register perception
g Word order indices
h Cross-linguistic paper on word order variation
i Workshop "Register and Word order"
j Cross-linguistic paper on register
k New outlook

Figure A06-2: Work plan identifying Work Packages 1–5 and resulting milestones a–k

Cooperation with external partners. We plan to invite Gregory Guy from the NYU Department of Linguistics, a specialist on cross-linguistic and cross-cultural perspectives in sociolinguistics, during the second year for 2 weeks in order to collaborate on variationist methodology and on the relation between register variation and change in progress. Furthermore, we plan to cooperate during our fieldwork (and beyond) with local linguists, in particular with the sociolinguist Yahya Modarresi from the Institute for Humanities and Cultural Studies in Tehran and with the anthropologist and linguist Barbara Blaha Pfeiler, CEPHCIS, UNAM Merida.

In the **second funding phase**, we plan to further study the social meaning of variants between registers, in order to better understand the reasons behind register choice and register change by speakers, across languages. Beyond that the focus of the second funding period will be on deepening our understanding of the dependency and interrelation of register and context in the choice between syntactic variants. The results of the second funding phase will help us to work in the third funding period on modeling a more comprehensive approach of register variation in grammar.

Bibliography

Adli, A. (Forthcoming). Topic chains in dialogues. *Journal of Pragmatics* (Prominence in Pragmatics).
– (2011). Gradient acceptability and frequency effects in information structure: A quantitative study on Spanish, Catalan, and Persian. Habilitation. Universität Freiburg.
– (2013). Syntactic variation in French wh-questions: A quantitative study from the angle of Bourdieu's sociocultural theory. *Linguistics* 51.3, 473–515.

– (2015). What you like is not what you do: Acceptability and frequency in syntactic variation. In: *Variation in Language: Usage-based Vs. System-based Approaches*. Ed. by A. Adli, M. G. García & G. Kaufmann. de Gruyter Mouton, 173–199.
– ed. (2016). *sgs corpus*. Sociolinguistic Lab at the University of Cologne. URL: http://sgscorpus.com (visited on 04/21/2019).
– (2017). Variation in style: Register and lifestyle in Parisian French. In: *Selected Papers from the 8th International Conference on Language Variation in Europe (ICLaVE 8)*. Ed. by I. Buchstaller & B. Siebenhaar. John Benjamins, 157–171.
Adli, A., E. Engel, L. Romary & F. Same (2018). A stand-off XML-TEI representation of reference annotation. Poster presented at the 40th Annual Conference of the German Linguistic Society (DGfS 2018), Stuttgart. URL: https://hal.inria.fr/hal-01876327 (visited on 04/22/2019).
Auer, P. & Y. Maschler (2013). Discourse or grammar? VS patterns in spoken Hebrew and spoken German narratives. *Language Sciences* 37, 147–181.
Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge UP. DOI: 10.1017/cbo9780511519871.
– (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast* 14.1, 7–34.
Cameron, R. (1992). Pronominal and null subject variation in Spanish: Constraints, dialects, and functional compensation. PhD thesis. University of Pennsylvania.
Carroll, M. & M. Lambert (2003). Information structure in narratives and the role of grammaticalized knowledge. In: *Information Structure and the Dynamics of Language Acquisition*. Ed. by C. Dimroth & M. Starren. John Benjamins, 267–287.
Dimroth, C., C. Andorno, S. Benazzo & J. Verhagen (2010). Given claims about new topics. How Romance and Germanic speakers link changed and maintained information in narrative discourse. *Journal of Pragmatics* 42, 3328–3344.
Dufter, A. (2008). On explaining the rise of *c'est*-clefts in French. In: *The Paradox of Grammatical Change: Perspectives from Romance*. Ed. by U. Detges & R. Waltereit. Current Issues in Linguistic Theory. John Benjamins, 31–56.
Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics* 12.4, 453–476. DOI: 10.1111/j.1467-9841.2008.00374.x.
Ferguson, C. A. (1959). Diglossia. *Word* 15, 325–340.
Féry, C. (2015). Extraposition and prosodic monsters in German. In: *Explicit and Implicit Prosody in Sentence Processing*. Ed. by L. Frazier & E. Gibson. Springer, 11–37.
Frey, W. (2006). Contrast and movement to the German prefield. In: *The Architecture of Focus*. Ed. by V. Molnár & S. Winkler. de Gruyter Mouton, 235–264.
Frommer, P. R. (1981). Post-verbal phenomena in colloquial Persian syntax. PhD thesis. University of Southern California.
Givón, T. (1979). From discourse to syntax: Grammar as a processing strategy. In: *Discourse and syntax*. Ed. by T. Givón. Academic Press, 81–112.
– (2009). Introduction. In: *Syntactic Complexity: Diachrony, Acquisition, Neuro-cognition, Evolution*. Ed. by T. Givón & M. Shibatani. John Benjamins, 1–19.
Göksel, A. & C. Kerslake (2005). *Turkish: A Comprehensive Grammar*. Routledge.
Götze, M., T. Weskott, C. Endriss, I. Fiedler, S. Hinterwimmer, et al. (2007). Information structure. In: *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*. Ed. by S. Dipper, M. Götze & S. Skopeteas. Vol. 7. Interdisciplinary Studies on Information Structure (Working Papers of the SFB 632). Universitätsverlag Potsdam, 147–187.
Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge UP. DOI: 10.1017/cbo9780511616983.
Haeri, N. (1989). Overt and non-overt subjects in Persian. *IPrA Papers in Pragmatics* 3.1, 155–167.
Haider, H. (2010). *The Syntax of German*. Cambridge UP.
Halliday, M. A. K. (1979). Differences between spoken and written language. In: *Communication through reading: Proc. of Fourth Australian Reading Conference*. Ed. by G. Page, J. Elkins & B. O'Connor. Australian Reading Association, 37–52.
Hartmann, K. (2013). Prosodic constraints on extraposition in German. In: *Rightward Movement in a Comparative Perspective*. Ed. by G. Webelhuth, M. Sailer & H. Walker. John Benjamins, 439–472.

Hickmann, M. & H. Hendriks (1999). Cohesion and anaphora in children's narratives: A comparison of English, French, German, and Mandarin Chinese. *Journal of Child Language* 26, 419–452.
Institut für Deutsche Sprache (2018). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2018-II (Release vom 12.11.2018)*. URL: www.ids-mannheim.de/DeReKo.
Karimi, S. (2005). *A minimalist approach to scrambling: Evidence from Persian*. de Gruyter Mouton.
Kunz, K. & E. Lapshinova-Koltunski (2015). Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies* 14.1, 258–288.
Labov, W. (2001). *Principles of Linguistic Change, Volume 2: Social Factors*. Wiley-Blackwell.
Lambert, W. E. (1967). The social psychology of bilingualism. *Journal of Social Issues* 23, 91–109.
Levinson, S. C. (2006). On the human interaction engine. In: *Roots of Human Sociality: Culture, Cognition and Interaction*. Ed. by N. J. Enfield & S. C. Levinson. Berg, 39–69.
Maas, U. (2010). Literat und orat. Grundbegriffe der Analyse geschriebener und gesprochener Sprache. *Grazer Linguistische Studien* 73, 21–150.
Miller, J. & R. Weinert (1998). *Spontaneous Spoken Language: Syntax and Discourse*. Oxford UP.
Modarresi, Y. (1978). A Sociolinguistic Analysis of Modern Persian. PhD thesis. University of Kansas.
Neumann, S. (2013). *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. de Gruyter.
– (2014). Cross-linguistic register studies: Theoretical and methodological considerations. *Languages in Contrast* 14.1, 35–57. DOI: 10.1075/lic.14.1.03neu.
Paolillo, J. C. (2000). Formalizing formality: an analysis of register variation in Sinhala. *Journal of Linguistics* 36.2, 215–259.
Pawley, A. & F. H. Syder (1983). Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics* 7, 551–576.
Pfeiler, B. & L. Zámišová (2006). Bilingual education: Strategy for language maintenance or shift of Yucatec Maya? In: *Mexican Indigenous Languages at the Dawn of the Twenty-first Century*. Ed. by M. G. Hidalgo. de Gruyter Mouton, 294–313. DOI: 10.1515/9783110197679.3.294.
Rühlemann, C. (2006). Coming to terms with conversational grammar: 'dislocation' and 'dysfluency'. *International Journal of Corpus Linguistics* 11, 385–409.
Sankoff, D. (1988). Sociolinguistics and syntactic variation. In: *Language: The Socio-cultural Context*. Ed. by F. Newmeyer. Vol. 4. Linguistics: The Cambridge Survey. Cambridge UP, 140–161.
Sato, Y. & S. Karimi (2016). Subject-object asymmetries in Persian argument ellipsis and the anti-agreement theory. *Glossa: A Journal of General Linguistics* 1.1, 8, 1–31.
Skopeteas, S. & E. Verhoeven (2009). The interaction between topicalization and structural constraints: Evidence from Yucatec Maya. *The Linguistic Review* 26.2–3, 239–259.
Stutterheim, C. von & M. Carroll (2005). Subjektwahl und Topikkontinuität im Deutschen und Englischen. *LiLi* 139, 7–27.
Temme, A. & E. Verhoeven (2016). Verb class, case, and order: A cross-linguistic experiment on non-nominative experiencers. *Linguistics* 54.4, 769–814.
Trudgill, P. (1974). *The Social Differentiation of English in Norwich*. Cambridge UP.
– (2011). *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford UP.
Trutkowski, E. (2016). *Topic Drop and Null Subjects in German*. de Gruyter Mouton.
Verhoeven, E. (2010). Agentivity and stativity in experiencer verbs: Implications for a typology of verb classes. *Linguistic Typology* 14, 213–251.
– (2014). Thematic prominence and animacy asymmetries: Evidence from a crosslinguistic production study. *Lingua* 143, 129–161.
– (2015). Thematic asymmetries do matter! A corpus study of German word order. *Journal of Germanic Linguistics* 27.1, 45–104.
Verhoeven, E. & N. Lehmann (2018). Self-embedding and complexity in oral registers. *Glossa: A Journal of General Linguistics* 3.1, 93.
Verhoeven, E. & S. Skopeteas (2015). Licensing focus constructions in Yucatec Maya: An empirical study on the association with focus. *International Journal of American Linguistics* 81.1, 1–40.