CHAPTER 7

Frequential test of (S)OV as unmarked word order in Dutch and German clauses

A serendipitous corpus-linguistic experiment

Gerard Kempen and Karin Harbusch
Max Planck Institute for Psycholinguistics, Nijmegen & Cognitive Psychology Unit, Leiden University / Department of Computer Science, University of Koblenz-Landau

In a paper entitled "Against markedness (and what to replace it with)", Haspelmath argues "that the term 'markedness' is superfluous", and that frequency asymmetries often explain structural (un)markedness asymmetries (Haspelmath 2006). We investigate whether this argument applies to Object and Verb orders in main (VO, marked) and subordinate (OV, unmarked) clauses of spoken and written German and Dutch, using English (without VO/OV alternation) as control. Frequency counts from six treebanks (three languages, two output modalities) do not support Haspelmath's proposal. However, they reveal an unexpected phenomenon, most prominently in spoken Dutch and German: a small set of extremely high-frequent finite verbs with unspecific meanings populates main clauses much more densely than subordinate clauses. We suggest these verbs accelerate the start-up of grammatical encoding, thus facilitating sentence-initial output fluency.

Keywords: SOV, SVO, markedness, German, Dutch, verb frequency, sentence planning, fluency, corpus linguistics, psycholinguistics

1. Introduction

Ten years ago, Martin Haspelmath published an influential paper entitled "Against markedness (and what to replace it with)", arguing "that the term 'markedness' is superfluous", and that "[i]n a great many cases, frequency asymmetries can be shown to lead to a direct explanation of observed structural asymmetries" (2006: 25). In the present chapter, we investigate whether this argument applies to the linear order of Object and Verb in main and subordinate clauses of German and Dutch. As is well-known, main clauses of these languages are VO, whereas subordinate

DOI 10.1075/z.210.07kem

clauses are OV – with very few exceptions. According to the currently dominant view in the linguistic literature, OV is the unmarked, VO the marked linear order (e.g. Koster 1975; Haider 2010). Instigated by Haspelmath's plea for a possible role of frequency in markedness phenomena, and expedited by the availability of large syntactically parsed corpora ("treebanks"), we decided to investigate whether the frequency distribution of VO and OV orders in clauses of German and Dutch is eligible as substitute for the unmarkedness of OV.

Haspelmath (2006: 35–36), following Dryer (1995), distinguishes two operational definitions of the "distributional markedness" of one member of a set of competing constructions, for instance, a set of alternate linear orders. On the first definition, one linear order – the marked one – is selected under "specified conditions", and the unmarked one may *always* occur, irrespective of whether or not the specified conditions hold. On the second definition, the marked order is exclusively reserved for the linear order meeting the specified conditions, and the unmarked option is realised *elsewhere* – in all cases where those conditions do not hold. The second definition is applicable to the choice between VO and OV in German and Dutch: VO is obligatory in main clauses, OV in finite and nonfinite subordinate clauses. In itself, this rule does not determine which order is marked: if the property "main" is deemed the specified condition to be checked first, then VO is the marked order (and OV the default option). But why do we not view "subordinate" as the special property? Therefore, we will proceed on the assumption that the OV-as-unmarked-order can be motivated synchronically in terms of rule system properties (e.g. Koster's (1975) grammar complexity argument based on the position of particles of separable verbs), or diachronically in terms of language change and grammaticalisation (e.g. Haider 2010).

Neither Haspelmath nor Dryer address constituent orders in clauses of German and Dutch. Hence, we do not know whether they would support the hypothesis that OV should be more frequent than VO if it is the unmarked option. What both these authors do address is the fact that extraneous factors can interfere with this prediction, i.e. neutralise the frequency imbalance, or even reverse the imbalance, leading to an invalid mapping from frequency to (un)markedness. With these potential error sources in mind, we have decided that frequency counts of clause types in German and Dutch should be extended with similar counts in a "control language", i.e. a language that is similar to Dutch and German in all relevant respects but has no VO/OV alternation linked to clause type. A suitable candidate is English. In this manner, we cast the frequential test in the form of a quasi-experimental design enabling us to isolate frequency effects due to markedness from frequency effects due to any other difference between main and subordinate clauses.

In sum, based on the OV-as-unmarked-order hypothesis we expect that, in Dutch and German, the number of OV clauses in a corpus of spoken or written

texts is higher than the number of VO clauses (> 50 percent of all occurrences of a finite or nonfinite clause). However, the proportion of VO – i.e. main – clauses may rise above the fifty-fifty ratio due to factors working in opposite direction to the (un)markedness hypothesis. We can check this possibility by comparing the proportions obtained from Dutch and German text corpora on the one hand with the proportion of main and subordinate clauses in a comparable English corpus on the other. If the German and Dutch proportions of subordinate clauses turn out to be a minority, thus falsifying the hypothesis, we can resort to the weaker prediction that this minority is still *larger* than in English.

In the following, we not only describe the frequential test and its results (Sections 2 and 3) but also an unexpected data pattern that we believe is informative about early lexico-syntactic processes during spoken and written sentence production (Section 4).

2. Methodology

The data sources we had at our disposal were the six syntactically annotated corpora listed in Table 1 (see also Appendix A): three treebanks with spoken, three with written text. The spoken materials we have analysed consists of sentences extemporaneously produced in varied dialogue situations (face-to-face, telephone); the written texts originated from printed materials (journal and magazine articles, book fragments). Together, the six treebanks contain more than 440,000 sentences, comprising almost 800,000 clauses.

By "clause" we mean a word group headed by a verb of any type (full, auxiliary, copula, modal), and we assume that every verb (of any type) is head of one clause (of any category – finite, infinitival, participial, gerund). As a consequence of these definitions, numbers of clauses will closely approximate numbers of verbs ("one verb, one clause"). Following the "topological" approach to word order in German and Dutch (Drach 1937; Höhle 1986), we assume that the head verb of a clause can be placed either at the so-called "left bracket" ("verb-second" in modern terminology), or at the "right bracket" ("verb-final"). In main clauses, the head verb is placed at the left bracket; in clauses of any other type, the head verb goes to the right bracket. The canonical positions for direct and indirect objects (as well as for many types of adjuncts/modifiers) are in between these brackets. Therefore, all main clauses (including imperatives) are VO, no matter whether they are actually followed by objects or adjuncts, or not. Likewise, Dutch and German subordinate clauses are always OV irrespective of whether or not the clause they are heading includes any preverbal constituents. Given these definitions, a sentence like *Sie hat noch nicht geantwortet* 'She hasn't replied yet' will be analysed as consisting of two

clauses: a finite main VO clause and a nonfinite subordinate OV clause; and the subordinate version … *dass sie noch nicht geantwortet hat* 'that she hasn't replied yet' contains two OV clauses.

Due to Gapping and other elliptical processes, not every clause contains an explicit verb. We did not try to compensate for such "missing" verbs. Word groups without any verbform (e.g. many titles of newspaper articles, or dialogue turns in the form of NPs or PPs without embedded clauses) were not considered in the calculations. Furthermore, we counted as subordinate clauses (hence, in Dutch and German as OV): prenominal participles (e.g. *the hastily leaving guests, the severely wounded driver, cheering crowds*) and nominalised verbs, including English gerunds, as well as Dutch and German constructions functioning very much like the English progressive (*aan*+*het*+infinitive, e.g. *Ze is een artikel aan het schrijven*, *Sie ist ein Papier am schreiben* 'She is writing a paper'). If a German or Dutch subordinate clause had been annotated as embodying VO order, we counted it as VO. Examples are embedded root clauses such as in German *Sie dachte, er wäre verheirated* 'She thought he was married'. Figure 1 shows the distribution of the number of verbs (and clauses) *per sentence* in each treebank.

Table 1. Treebanks used in the present study (for details, see Appendix A). The first data column shows the number of trees containing at least one verbform, the second data column the number of verbform tokens used in the calculations (repaired/edited utterance fragments were discarded).

Language and modality		Treebank	Number of sentences	Number of verbforms
German	spoken	VM	38,328	50,676
	written	TIGER	50,474	106,912
Dutch	written	CGN	126,787	162,985
	spoken	LASSY	65,061	140,695
English	spoken	SWB	110,504	167,272
	written	WSJ	49,208	160,899

For all treebanks (except TIGER) we had to *lemmatise* the verbforms, i.e. to assign them to a citation form ("lemma"; the infinitive, except in case of English modal auxiliaries and a few defective verbs). A major subtask here concerned separable verbs: combining the particle with the core verb. For lemmatisation purposes, we used computational-linguistic software available in the literature or developed in-house, but we carefully checked the results manually. When reporting verb frequencies, we will always use the citation forms (lemmas). In order to obtain the "*total verb frequency*" of a verb, we added the frequencies of all its (inflected) forms. Excluded from all calculations were verbs within sentence fragments tagged as repairs or revisions (virtually restricted to the spoken corpora).

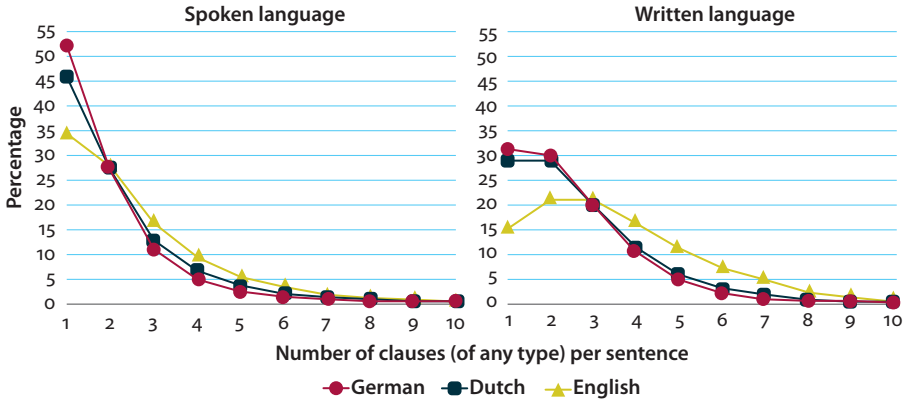

Figure 1. Percentage of sentences containing n ($1 \leq n \leq 10$) clauses (verbs) per sentence.

Importantly, we did not try to *disambiguate* verbforms. That is, if a verbform can be allocated to more than one infinitive (e.g. *lay* as finite form of *lie* or *lay*), we arbitrarily chose one (always the same). If the citation form itself is ambiguous, that is, belong to multiple subclasses of verbs (e.g. intransitive or transitive, full verb or auxiliary), we adopted the verb class tag already attached to the verbform in the treebank; we did not try to disambiguate polysemous or homophonous verbs (e.g. *lie*). In sum, we worked with the parse tree information stored in the treebanks as much as possible, deviating from it only in case of obvious parsing errors or lacunae.

As final preparatory step we assigned a "clause type" to each individual verbform token. We distinguished three types of clauses: "main" (including imperatives and parentheticals such as *you know*), "finite subordinate" (complement, adverbial, and relative clauses), and "nonfinite" (infinitival, participial, and gerund). (As convenient abbreviations we will use MAIN-FIN, SUB-FIN, and NONFIN.) For each of these three clause types, and for each treebank separately, we defined a set of search queries based on the treebank's morphological, lexical and syntactic tagging system and on the relative positions of these tags and other node labels in the syntactic trees.

Before turning to the frequency test we need to introduce two crucial verb parameters derived from verb tokens in each of the corpora: "bias" and "coverage". Both parameters are explained and illustrated in Appendix B. In order to obtain the values of these parameters, one first computes, in each corpus, the rank order of all verb lemmas with respect to their *total* frequency of occurrence. Table 2 (rightmost column) shows the frequencies of the seven highest-frequency verbs in SWB ("Top7"). The three middle columns show the distribution of the occurrences over the three clause types we distinguish. The "bias" of a verb with respect to a clause type is defined as the number of its occurrences as percentage of its total

frequency. The "coverages" of a verb vis-à-vis the various clause types are calculated as percentages of the total number of verb tokens heading that clause type *in a larger group of verbs*, e.g. in the entire corpus.

Table 2. The seven top-frequent verbs in the SWB treebank. The numbers represent verbform tokens.

Lemma	MAIN-FIN	SUB-FIN	NONFIN	TOTAL
go	876	646	3,722	5,244
think	3,604	436	1,368	5,408
get	1,196	1,028	3,412	5,636
know	10,184	365	1,859	12,408
do	7,581	3,319	3,337	14,237
have	7,337	4,353	3,674	15,364
be	22,164	13,353	5,387	40,904
Top7 verbs	**52,942**	**23,500**	**22,759**	**99,201**
ALL VERBS	**75,475**	**36,913**	**54,884**	**167,272**

3. Three frequential tests

If the unmarkedness of OV order in Dutch and German subordinate clauses is reflected frequentially, then the percentage of subordinate clauses within the total number of clauses should be higher than the corresponding percentage of main clauses. The rightmost column of Table 3 shows that this holds for written texts in these languages, but not for spoken texts. Although this looks like partial confirmation of the hypothesis, the last two numbers in the column show that the analogous percentages for English texts are in the same range. Obviously, this data pattern fails to support the hypothesis. Moreover, considering finite clauses only in the first two data columns, we see that the MAIN-FIN to SUB-FIN ratios in all corpora are opposite to the prediction, in Dutch and German more so than in English.

However, this conclusion may be too hasty. Proponents of the "unmarked = more frequent" hypothesis might argue that we could just as well analyse the corpus data at the level of individual verbs, i.e. determine the biases vis-à-vis the clause types "unweighted" for the total frequency of the verbs. We did these calculations, and the results are summarised in Table 4, which shows the mean subordinate-clause bias of all verbs (SUB-FIN or NONFIN), unweighted for frequency. (The weighted frequencies are in the rightmost column of Table 3.) Although, in Dutch and German, the unweighted percentages are all above 50, in line with the hypothesis, the impact of this finding is annulled by the corresponding English percentages where the same bias shows up.

Table 3. Relative frequency of three types of clauses in the six treebanks. The first numerical data column denotes main clauses (VO in German and Dutch), the last column all subordinates (OV in German and Dutch). The numbers are percentages. Corresponding percentages in the first and the last data column add up to 100.

Treebank	Main, (MAIN-FIN)	Finite subordinate (SUB-FIN)	Nonfinite (NONFIN)	All subordinate
German Spoken	68.6	8.7	22.7	31.4
German Written	45.5	19.2	35.3	54.5
Dutch Spoken	56.1	16.1	27.8	43.9
Dutch Written	44.1	17.1	38.7	55.9
English Spoken	45.1	22.1	32.8	54.9
English Written	32.0	24.6	43.4	68.0

Table 4. Average subordinate clause biases (second data column), unweighted for total frequency. The first data column shows the number of verbs (citation forms) in each treebank.

Treebank and modality		Total number of different verbs (lemmas)	Subordinate-clause bias (SUB-FIN + NONFIN) Unweighted means
German	Spoken	1,083	71.7
	Written	4,892	71.2
Dutch	Spoken	3,884	79.6
	Written	4,364	76.1
English	Spoken	2,564	81.8
	Written	4,082	82.9

In a final attempt to rescue the hypothesis, we focused on frequency *differences* between verbs. We speculated that the hypothesis might hold only for verbs occurring with low or intermediate frequency, but not for high-frequency verbs. The Dutch and German pattern revealed by Tables 3 and 4 might be due to an overwhelming proportion of main clauses headed by high-frequent verbs, obliterating a bias in favor of OV in low- and mid-frequent verbs. Therefore, we decided to do the following calculations, separately for the six corpora: the rank order of all verb lemmas with respect to their *total* corpus frequency; and, for each verb lemma, its bias and coverage percentages with respect to the three clause types, *weighted* for frequency.

As expected, a small number of very high-frequent verbs covers a huge proportion of verb occurrences. To give an impression, in Figure 2 we show the coverage of the 50 verbs with highest total frequency in their corpus (henceforth called the "Top50" of that corpus; the number 50 is arbitrary). The chart reveals that the Dutch and German Top50 verbs have very similar coverage percentages, and that their

English counterparts tend to be a little higher. This latter tendency runs counter the "unmarked = more frequent" hypothesis, which expects the two types of subordinate clauses (labelled SUB-FIN and NONFIN) to have higher coverages in Dutch and German where these clause types embody the unmarked OV word order.

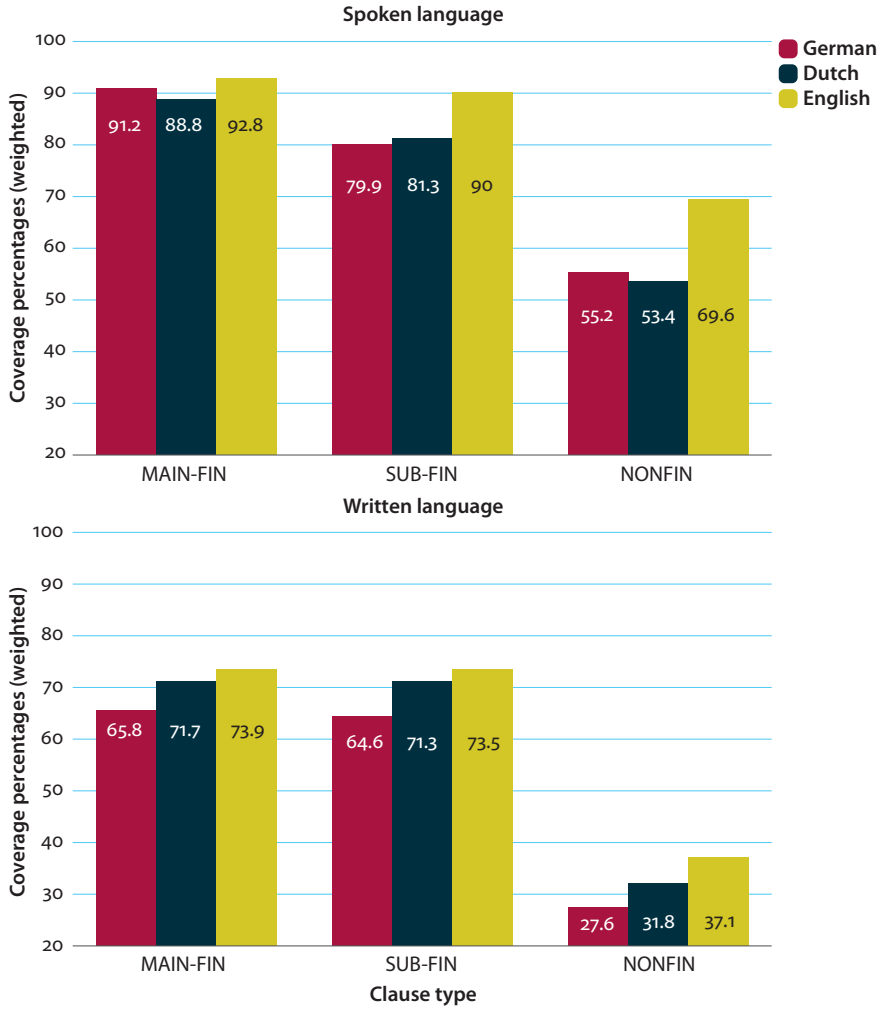

Figure 2. Coverage of Main, Finite Subordinate, and Nonfinite clauses by high-frequency verbs (Top50) in each of the six treebanks. Percentages weighted for frequency.

In order to test whether less frequent verbs have different bias patterns vis-à-vis the three clause types, we calculated these separately for the hapax legomena ("low-frequent"), for the verbs with intermediate frequencies (i.e. from 2 to the maximum below the Top50 frequencies ("mid-frequent"), and for the Top50 verbs

("high-frequent"). For the verbs in each of these frequency ranges, we computed average biases with respect to the three clause types, weighted for frequency.[1]

Figure 3 indeed reveals considerable cross-frequency variation in bias patterns. Nonfinite biases are preponderant in low-frequent verbs in all treebanks, and there is a bias shift from nonfinite to main-clause with increasing verb frequency.

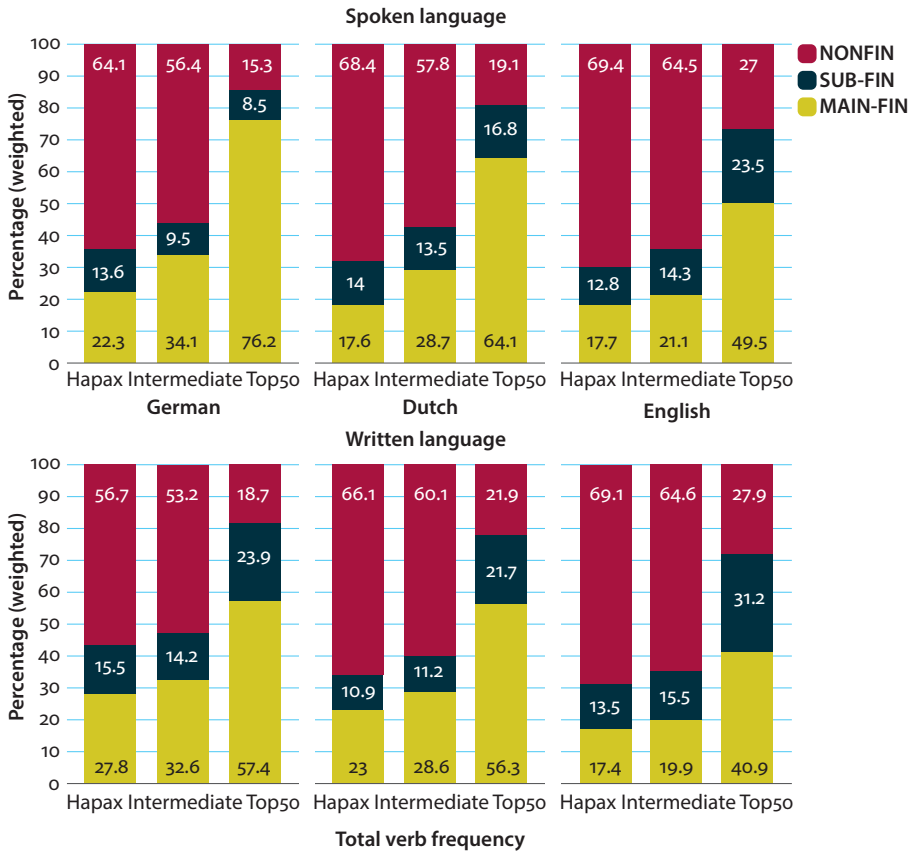

Figure 3. Average MAIN-FIN, SUB-FIN and NONFIN biases (weighted for frequency) as a function of total verb frequency (hapax, intermediate, Top50). The colors in each bar represent the distribution of verbform tokens across clause types (yellow/bottom: NONFIN; black/middle: SUB-FIN; red/top: NONFIN). Notice that, although all bars are equally high (all adding up to 100), they represent widely differing numbers of verb tokens (coverages).

1. To illustrate this for the 1015 mid-frequent verbs (lemmas) in SWB: in this frequency range, the treebank contains 5,232 verbform tokens in a main clause, 3,553 tokens in a finite subordinate clause, and 15,984 tokens in a nonfinite clause, totaling 24,769 verbforms. The mean main-clause bias of the mid-frequent verbs, is given by 5,232/24,769*100=21.1%. This number is one of the percentages shown in the top-right group of bars in Figure 3.

However, the key hypothesis is not supported, due to the fact that the English treebanks show the same "main-clause bias shift" as the Dutch and German ones. Figure 4 shows that, if the mean bias percentages are not weighted for frequency, the same pattern emerges, although somewhat less clear-cut. What the unweighted means bring out more saliently than the weighted ones, is the restriction of the main-clause bias shift to a small set of high-frequent verbs – presumably not many more than 50 verbs (Appendix C).

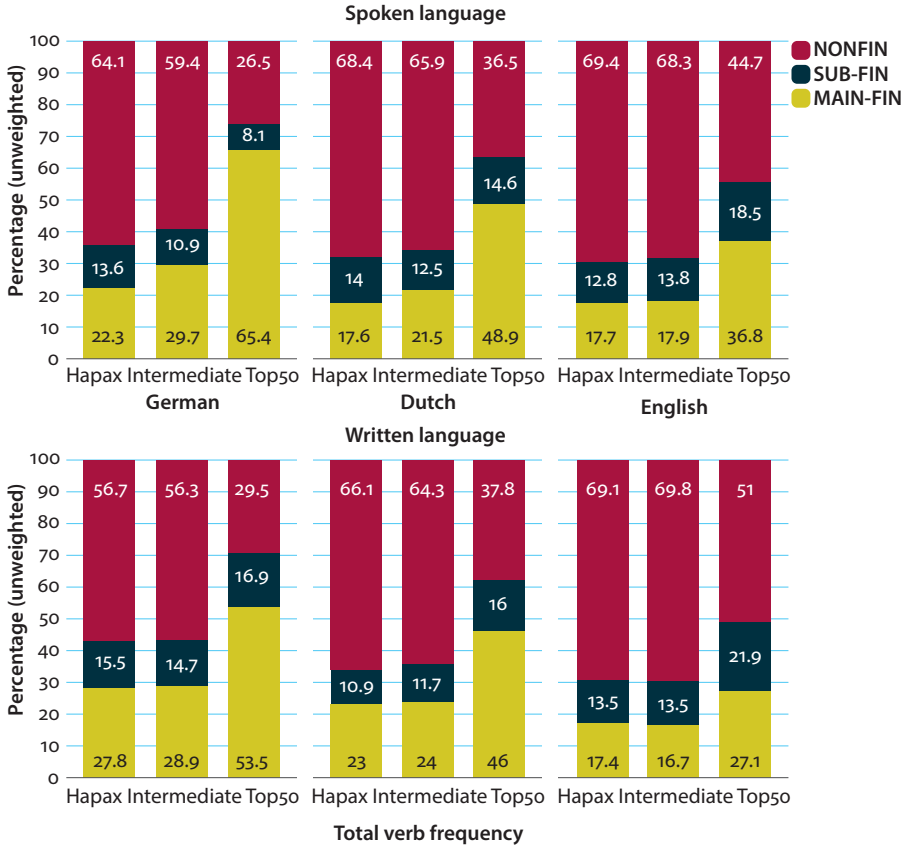

Figure 4. Average MAIN-FIN, SUB-FIN and NONFIN biases (unweighted for frequency) as a function of overall verb frequency (hapax, intermediate, Top50). Based on the same data as Figure 3, which displays biases weighted for frequency.

In conclusion, none of the three data explorations reported in this section confirmed the predictions derivable from the hypothesis that the unmarked OV word order in Dutch and German subordinate clauses should be reflected by a higher ratio of the number of subordinate clauses to the number of main clauses (VO),

compared to the corresponding ratio in English, where VO is common to all clause types. Hence, the assumption of OV as unmarked word order in clauses of present-day German and Dutch cannot be founded on frequential evidence of the type reported here.

4. Discussion: Time and fluency pressures can boost VO:OV ratios

What could have caused the high incidence of VO structures and the failure of the frequential test? In this section we argue that high-frequent verbs at relatively early position (VO instead of OV) in the clause they are heading act as powerful fluency facilitators at sentence onset.

The observed data pattern can be characterised in statistical terms as one main effect and two interactions. The main effect is the overrepresentation of a small set of high-frequent verbs in main clauses compared to finite subordinate clauses – the *main-clause bias shift*. This effect interacts with modality, being more salient in spoken than in written language, and with target language – being stronger in German and Dutch than in English.[2]

A first clue to explain the observed effects can be gleaned from properties of the verbs the Top50s are composed of – see Appendix C. It is unsurprising that, in all corpora, the Top50s include verbs denoting frequent events and states-of-affairs in the content domain under discussion. But the Top50s are more densely populated by domain-independent verbs with *functional/pragmatic* meanings: modal verbs, verbs of communication and cognition, verbs expressing the producer's propositional attitude, evidential verbs. Also prominent are light verbs, copulas, and auxiliaries – three types of verbs whose *raison d'être* is syntactic rather than semantic. Hence, with the exception of a few domain-specific verbs, one can characterise the Top50 members as general-purpose verbs with unspecific meanings.

Why would verbs of these categories attract main clauses – or be attracted to main clauses (for we don't know the direction of causation) – thereby engendering a main-clause bias shift? One hypothesis coming to mind concerns the functional/pragmatic meanings expressed by Top50 members. One could argue, in particular with respect to evidential and propositional attitude verbs, that such meanings need to be expressed only once per sentence/proposition, hence most naturally in main clauses. However, the observed interaction between main-clause bias shift and modality then would force us to assume that the language user's need to express

2. The data suggest that the main-clause shift affects German more strongly than Dutch. We tentatively account for this contrast in terms of competition between (near-)synonymous lexico-syntactic structures (see end of present section).

such meanings is less pressing in written than in spoken sentences – an assumption that may or may not be true. Another problematic issue raised by this hypothesis is that it does not fare well with the cross-linguistic interaction: It would entail that speakers of English express functional/pragmatic meaning aspects less readily than speakers of German and Dutch.

We prefer an account that takes the observed interaction of main-clause bias shift with modality as point of departure. It is hardly controversial, at least regarding the three languages targeted in the present study, that the grammatical encoding process for a pluriclausal sentence tends to proceed hierarchically. It begins with the main clause (probably even when a finite complement or adverbial clause takes sentence-initial position) and, within this process, a head, *in casu* a finite verb, is selected at an early stage. Even less controversial is the assumption that high-frequent lexical items can be activated and retrieved from the mental lexicon more easily and rapidly than less frequent items. Furthermore, speakers are often under time pressure, attempting to generate the upcoming dialogue turn without empty pauses before utterance onset. Taken together, these considerations suggest that, at least in the absence of editing and repair opportunities, speakers can benefit from having at their disposal a small collection of easily accessible (highly available) verbs that allow a quick start-up of the sentence and a fair chance of completing it grammatically. Thus, the sentence can get going fluently at an early point in time, allowing the speaker some extra time and processing capacity to formulate the real content of the sentence – the proposition or message referring to events or states-of-affairs in the world. Consequently, the conceptual properties carried by spoken main-clause head verbs will be relatively domain-neutral. Given this scenario, time and fluency pressures are expected to be milder in finite subordinate clauses. Activation and retrieval of domain-specific verbs can take place partly in parallel with activation of the lexical materials for the main clause, partly in parallel with overt phonetic realisation of the main clause. Hence, in finite subordinate clauses, topical verbs have extra time to build up activation and a better chance to win the head-of-clause competition.

The account so far covers two out of the three reported effects: the main-clause bias shift and, because time and fluency pressures are usually weaker while writing than while speaking, the interaction with modality. Can it also explain the interaction with target language (stronger main-clause bias shift in German and Dutch than in English)?

The following reasoning leads to an answer in the affirmative. It capitalises on the fact that, in English, main and subordinate clauses are both VO. Recent empirical and computational-modelling work on sentence production emphasises that the difficulty of producing a given syntactic structure is not always due to properties of the structure itself but to competition with other structures that can express the same conceptual content, especially when there are frequently used alternatives (Fitz, Chang & Christiansen 2011; MacDonald, Montag & Gennari 2016). Consider

the options available to speakers who are planning to revise a subordinate clause. When speaking Dutch and German, they are torn between maintaining the current OV structure, or switching to a much more frequent VO structure. The latter option causes the conceptual message that was originally planned as a subordinate clause, to be realised as the main clause of a *new* sentence.[3] Scenarios of this type – VO reformulations of OV clauses that were "nipped in the bud" – increase the main-clause bias shift. The absence of VO vs. OV competition benefits speakers of English compared to speakers of Dutch and German: The greater overall similarity of building plans for main and subordinate clauses often obviates the need to abandon the current clause plan. This predicts that the ratio of MAIN-FIN bias to SUB-FIN bias is smaller in English than in Dutch and German, as verified in Figures 3 and 4.

Factors related to variability of the building plans for clauses also yield an explanation for the difference between German and Dutch: German word order is more variable than Dutch word order, hence the competition between structural alternates may be fiercer.

The argument developed here enables a positive answer to the question posed in the title of the present section, and identifies a potent factor underlying the high incidence of VO relative to OV structures: facilitation of fluency at sentence onset during speaking. The well-known tendency for language users to mirror perceived frequency patterns in their own language output creates a positive feedback loop that will boost VO-to-OV ratios in Dutch and German even further. However, we hasten to add that this cannot be the entire story: the existence of strictly (S)OV languages like Japanese and Korean entails that additional constraints must be involved. To be continued.

References

Beek, Leonoor van der, Gosse Bouma, Robert Malouf & Gertjan van Noord. 2002. The Alpino Dependency Treebank. In Tanja Gaustad (ed.), *Computational Linguistics in the Netherlands 2001*. Amsterdam: Rodopi.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith & Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation* 2. 597–620. doi:10.1007/s11168-004-7431-3

Charniak, Eugene, Don Blaheta, Niyu Ge, Keith Hall, John Hale & Mark Johnson. 2000. *BLLIP 1987–89 WSJ Corpus Release 1 LDC2000T43. DVD*. Philadelphia: Linguistic Data Consortium.

3. That such OV-to-VO clause revisions – covert or overt – in spoken language are by no means rare, we demonstrated in a case study about the incidence of VO order after the German subordinating conjunction *weil* 'because' (Kempen & Harbusch 2016).

Drach, Erich. 1937. *Grundgedanken der deutschen Satzlehre*. Frankfurt am Main: Diesterweg. [Reprinted in 1963]

Dryer, Matthew. 1995. Frequency and pragmatically unmarked word order. In Mickey Noonan & Pamela Downing (eds.), *Word order in discourse*, 105–135. Amsterdam: John Benjamins. doi:10.1075/tsl.30.06dry

Eerten, Laura van. 2007. Over het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, 12. 194–215.

Fitz, Hartmut, Franklin Chang & Morten H. Christiansen. 2011. A connectionist account of the acquisition and processing of relative clauses. In Evan Kidd (ed.), *The acquisition of relative clauses. Processing, typology and function*, 39–60. Amsterdam: Benjamins. doi:10.1075/tilar.8.04fit

Godfrey, John J., Eduard C. Holliman & Jane McDaniel. 1992. SWITCH-BOARD: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Audio, Speech and Signal Processing (ICASSP-92)*, 517–520.

Haspelmath, Martin. 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42. 25–70. doi:10.1017/S0022226705003683

Haider, Hubert. 2010. Wie wurde Deutsch OV? Zur diachronen Dynamik eines Strukturparameters der germanischen Sprachen. In Arne Ziegler (ed.), *Historische Textgrammatik und Historische Syntax des Deutschen – Traditionen, Innovationen, Perspektiven*, 11–32. Berlin: De Gruyter. doi:10.1515/9783110219944.9

Höhle, Tilman N. 1986. Der Begriff 'Mittelfeld': Anmerkungen über die Theorie der topologischen Felder. In Walter Weiss, Herbert E. Wiegand & Marga Reis (eds.), *Akten des VII. Internationalen Germanistenkongresses*, 329–340. Tübingen: Niemeyer.

Hoekstra, Heleen, Michael Moortgat, Ineke Schuurman & Ton van der Wouden. 2001. Syntactic annotation for the spoken Dutch corpus project (CGN). *Language and Computers* 37(1). 73–87.

Kempen, Gerard & Karin Harbusch. 2016. Verb-second word order after German *weil* 'because': Psycholinguistic theory from corpus-linguistic data. *Glossa: a journal of general linguistics* 1(1). 1–32. doi:http://dx.doi.org/10.5334/gjgl.46

König, Esther & Wolfgang Lezius. 2003. *The TIGER language: A Description Language for Syntax Graphs, Formal Definition*. Stuttgart: University of Stuttgart.

Koster, Jan. 1975. Dutch as an SOV Language. *Linguistic analysis* 1. 111–136.

MacDonald, Maryellen C., Jessica L. Montag & Silvia P. Gennari. 2016. Are there really syntactic complexity effects in sentence production? A reply to Scontras, (2015). *Cognitive Science*, 40. 513–518.

Noord, Gertjan van, Gosse Bouma, Frank van Eynde, Daniël de Kok, Jelmar van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, & Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns & Jan Odijk (eds.), *Essential Speech and Language Technology for Dutch*, 147–164. Springer, Berlin. doi:10.1007/978-3-642-30910-6_9

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste & Ineke Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for Dutch*, 219–247. Berlin: Springer. doi:10.1007/978-3-642-30910-6_13

Stegmann, Rosemary, Heike Telljohann & Erhard W. Hinrichs. 2000. *Stylebook for the German Treebank in Verbmobil*. Saarbrücken: DFKI Report 239.

Wahlster, Wolfgang (ed.). 2000. *Verbmobil: Foundations of speech-to-speech translation*. Berlin: Springer. doi:http://dx.doi.org/10.1007/978-3-662-04230-4

Appendices

Appendix A. The six treebanks used in the present study

Table A1. Some important details concerning the treebanks.

Language and modality		Full name of treebank and key references	Abbreviated name
German	spoken	*VERBMOBIL Corpus* Stegmann, Telljohann & Hinrichs (2000); Wahlster (2000)	VM
German	written	*TIGER Corpus* Brants et al. (2004)	TIGER
Dutch	spoken	*Corpus Gesproken Nederlands 2.0* Hoekstra et al. (2001); Van Eerten (2007)	CGN
Dutch	written	*LASSY-Small Corpus* Oostdijk et al. (2013); Van Noord et al. (2013)	LASSY
English	spoken	*SWITCHBOARD Corpus* Godfrey, Holliman & McDaniel (1992)	SWB
English	written	*Wall Street Journal Corpus* Charniak et al. (2000)	WSJ

From the spoken VERBMOBIL dialogues, we used the sentences syntactically annotated in the TüBa-D/S treebank (Stegmann et al., 2000). The TIGER treebank of German consists of newspaper texts. These treebanks specify the same part-of-speech (PoS) tags for verbs. However, the encodings of sentence structure differ considerably. TüBa-D/S specifies *topological fields*, which allow easy classification of clauses as MAIN-FIN vs. SUB-FIN. TIGER does not specify clause type, but NONFIN can be identified easily in terms of PoS tags. For the other clause types, we used queries in TIGERSearch (König & Lezius 2003) along with JAVA programs we developed ourselves. A disadvantage of TIGER and VERBMOBIL compared to the Dutch and English corpora concerns present participles, which are all labeled as adjectives. In order to maintain the comparability of German and Dutch counts, we disregarded present participles in CGN and LASSY. We did *not* discard English present participles because they had received the same tag as gerunds (both suffixed with *-ing*) in the treebanks; moreover, quite a few verbs ending in *-ing* had been tagged there as adjectives or nouns. We estimate that, due to these between-treebank annotation differences, the numbers of NONFINs in the German and Dutch treebanks are somewhat too low relative to the English treebanks. However, we are confident that none of the conclusions drawn from the results in Section 3 are weakened by the procedure we followed.

CGN contains spoken sentences from various different domains (news, telephone conversations, speeches, etc.). However, not all of them were produced spontaneously. In total, we discarded about 3,800 sentences with read speech. LASSY contains written texts from a great variety of sources – not only newspaper articles but also excerpts from books, manuals, legal texts, the Dutch Wikipedia, etc. In the two Dutch treebanks, the sentences had been annotated with the same, relatively theory-neutral dependency graphs (Hoekstra et al. 2001; Van der Beek et al. 2002). Both corpora specify features that directly allow classifying clauses as MAIN-FIN and SUB-FIN. The LASSY treebank was queried in part through DACT (Van Noord et al. 2013), CGN through TIGERSearch; in both cases, we supplemented the queries with JAVA programs of our own making.

SWB is a large corpus of conversational dialogues comprising about 2,500 phone conversations by 500 speakers from around the USA.

WSJ contains three years of text from the Wall Street Journal (the ACL/DCI corpus). SWB and WSJ use very similar annotations. They do not specify features enabling straightforward identification of MAIN-FIN and SUB-FIN clauses. We rectified this by adapting TIGERSearch and/or writing our own JAVA software.

For more details, please contact the second author.

Appendix B. Computing bias and coverage values

VERB BIAS. Consider the verb *go* in Table 2. This verb has a strong bias (=(3,722/5,244)*100=70.0%) in favour of being head of a nonfinite clause (due to *going*/*gonna* playing a role in the progressive), and weaker SUB-FIN (12.3%) and MAIN-FIN (16.7%) biases. *Be*, on the other hand, has a strong bias toward heading main clauses (54.2%). Importantly, the *biases* of a verb are computed as a percentage of its *own* total frequency. In order to calculate the average bias of a *group* of verbs vis-à-vis a clause type, one can proceed in either of two ways. One can sum the raw numbers underlying the individual biases (as is done in row "Top7 verbs"), and divide by the summed total frequencies of the verbs in the group. For instance, the Top7 verbs together have a bias of 53.4% (= (52,942/99,201)*100) in favour of MAIN-FIN. This average is weighted for frequency of the underlying individual verb biases. In the second procedure, one first computes the bias percentages for each verb separately, followed by adding and averaging these percentages. This resulting mean value is *unweighted* for frequency.

VERB COVERAGE. Consider the bottom row of Table 2. Of the 75,475 main clauses in SWB, 22,164 were a form of *be*, yielding a coverage of 29.4% in the Top7. The *total* coverage of *be* in the collection of all SWB verbs is 40,904 of the 167,274 clauses, i.e. 24.5%. We also report coverage percentages for certain verb *groups*, e.g. for all verbs sharing some property, e.g. belonging to the Top7, or being a "hapax legomenon" (i.e. being a verb whose total frequency in the corpus equals 1). For example, the MAIN-FIN coverage of the Top7 in the entire SWB corpus is no less than 70.1% (=52,942/75,475)*100), although these 7 verbs comprise only 0,3% of the 2564 different verbs (lemmas) appearing in the corpus. Notice that this coverage value is weighted for frequency of the individual group members. One obtains an unweighted coverage percentage by calculating coverage percentages per verb, followed by adding and averaging these percentages.

Appendix C. The **Top50 verbs (lemmas) in the six treebanks**

The verbs are printed in ascending order of total lemma frequency.

VERBMOBIL (spoken German)
reichen, ausschauen, ankommen, freihaben, anhören, vereinbaren, freuen, losfahren, festhalten, halten, reservieren, bleiben, dauern, kümmern, heißen, tun, meinen, kosten, grüßen, liegen, ausmachen, schauen, lassen, glauben, kommen, finden, brauchen, buchen, mögen, vorschlagen, fliegen, aussehen, treffen, sehen, wissen, geben, nehmen, denken, passen, wollen, sollen, fahren, sagen, machen, müssen, gehen, werden, können, haben, sein

TIGER (written German)
scheinen, entscheiden, übernehmen, ablehnen, bestehen, ankündigen, mitteilen, setzen, erwarten, tun, erreichen, beginnen, meinen, gehören, schaffen, sprechen, mögen, nehmen, berichten, erhalten, nennen, wissen, stellen, finden, führen, fordern, gelten, bringen, zeigen, erklären, halten, heißen, liegen, bleiben, dürfen, sehen, gehen, stehen, kommen, lassen, machen, geben, sagen, wollen, müssen, sollen, können, haben, werden, sein

CGN (spoken Dutch)
heten, schrijven, bellen, spreken, geloven, eten, kopen, halen, praten, vertellen, proberen, lijken, nemen, zetten, spelen, kennen, gebeuren, lopen, lezen, houden, vragen, bedoelen, beginnen, blijven, werken, liggen, horen, laten, geven, mogen, krijgen, maken, kijken, staan, zien, willen, zitten, vinden, komen, denken, weten, worden, zullen, doen, zeggen, moeten, kunnen, gaan, hebben, zijn

LASSY (written Dutch)
vragen, zorgen, raken, spreken, ontstaan, denken, aflopen, vormen, zetten, lijken, voorkomen, spelen, gebeuren, volgen, noemen, vallen, bepalen, leiden, zitten, bestaan, stellen, gebruiken, brengen, werken, beginnen, liggen, weten, blijken, mogen, nemen, houden, laten, blijven, vinden, zeggen, geven, staan, doen, zien, krijgen, willen, maken, komen, gaan, moeten, zullen, kunnen, hebben, worden, zijn

SWB (spoken English)
love, believe, stay, remember, sound, spend, play, happen, enjoy, give, call, watch, keep, tell, let, buy, read, find, may/might, hear, pay, feel, put, seem, need, live, start, look, talk, try, use, come, work, make, want, take, like, guess, say, see, mean, can/could, shall/should/will/would, go, think, get, know, do, have, be

WSJ (written English)
receive, work, base, raise, try, provide, show, lead, decline, agree, find, remain, become, accord, know, end, help, offer, increase, call, begin, close, hold, think, want, see, give, pay, report, come, add, continue, fall, use, buy, go, get, take, rise, include, sell, expect, may/might, make, can/could, do, shall/should/will/would, have, say, be