Bernhard Wälchli*

The interplay of contrast markers (‘but’), selectives (“topic markers”) and word order in the fuzzy oppositional contrast domain

https://doi.org/10.1515/lingty-2022-0019
Received April 18, 2022; accepted March 15, 2023; published online April 20, 2023

Abstract: This investigation is a large-scale comparative corpus study of the oppositional contrast domain (also called “semantic opposition”) based on parallel texts. Oppositional contrast is established as a fuzzy region of the similarity space of contrast (‘but’), a domain also characterized by the occurrence of selectives (“topic markers”) and of initial non-predicative phrases in VSO/VOS-languages. Major findings are that many languages have special oppositional contrast markers and that there is a continuum between oppositional contrast markers and selectives, although truly intermediate markers are rare. The gradualness between oppositional and counterexpectative contrast is explained by semantic fuzziness and by emphasis, with strong emphasis being dependent on scales. Contrast is a rhetorical discourse relation and strong oppositional contrast can be used as a persuasive strategy aiming at establishing new common ground stepwise. The fuzziness of oppositional contrast has major theoretical and methodological implications. The encoding of the domain neither follows strict universals nor is it maximally diverse (diversity is strongly constrained). Due to its syntactic properties, oppositional contrast cannot be conceived of merely as a preestablished extralinguistic semantic domain. Furthermore, contrast exhibits a high degree of language-internal variability. General trends are reflected both by stable and by emergent grammar.

Keywords: contrast; emphasis; parallel texts; semantic maps; topic markers; word order

1 Introduction

This study explores the interplay of contrast markers, selectives – also known as “topic markers” – and non-dominant word order in a massive parallel text corpus:

*Corresponding author: Bernhard Wälchli [ˈbærnhaːrd ˈvaʊ̯χli], Institutionen för lingvistik, Stockholms universitet, SE – 106 91 Stockholm, Sweden, E-mail: bernhard@ling.su.se

© Open Access. © 2023 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.
translations of the New Testament.\textsuperscript{1} This section starts by introducing contrast. Selectives and non-dominant word order are treated further down.

In earlier literature, three different subdomains of contrast have often been distinguished, which are illustrated by (1)–(3). The terminology of Mauri (2008) is adopted.\textsuperscript{2} I will call the first sentence of the contrast construction \textsc{anchor sentence} $[A]$ and the second one \textsc{contrast sentence} $[C]$. “Sentence” rather than “clause” is used because contrasted sentences can contain subordinate clauses.

(1) \textsc{oppositive contrast}

$[[\text{John}]_{\text{AP}} \text{is tall}]_{A}, \text{but } [[\text{Bill}]_{\text{OP}} \text{is short}]_{C}$.

(2) \textsc{counterexpectative contrast}

$[\text{John is tall}]_{A}, \text{but } [\text{he's no good at basketball}]_{C}$.

(3) \textsc{corrective contrast}

$[\text{Pat is not a dentist}]_{A}, \text{but } [\text{a linguist}]_{C}$.

As far as contrast is concerned, this paper concentrates on oppositive contrast, which, however, for a better understanding of the notion, is situated within contrast more generally. Oppositive contrast not only opposes two sentences, but also two phrases (underlined in (1)). The \textsc{contrasted phrase} in the anchor sentence will be called \textsc{anchor phrase} $[\text{AP}]$ and the contrasted phrase in the contrast sentence \textsc{opposite phrase} $[\text{OP}]$. A major point made in this article is that the syntactic unit oppositive phrase is crucial for oppositive contrast, which entails that oppositive contrast is not only a meaning, but also has specific syntactic properties.

Oppositive contrast is semantically and pragmatically gradual, an effect that is mainly due to the rhetorical character of contrast, which is context-dependent (absent in artificial examples such as (1)). Examples (4a–c) represent increasingly rhetorically stronger opposition of two referents. Semantically, stronger

\footnotesize
\begin{itemize}
\item The source of inspiration is Sæbø (2002), who controversially claims that the German contrast marker \textit{aber} is a topic particle.
\item Alternative terms such as those in (i) have been used (for more terms, see Rudolph 1996: 172–174).
\end{itemize}

\small

(i) a. oppositive contrast: semantic opposition (Lakoff 1971), parallel contrast (Dik 1997), contrast (Malchukov 2004; Myhill and Xing 1996), double contrast (Umbach 2005)\hfill
\item b. counterexpectative contrast: denial of expectation (e.g. Lakoff 1971), adversion (Malchukov 2004), simple contrast (Umbach 2005)\hfill
\item c. corrective contrast: correction (Malchukov 2004; Umbach 2005)\hfill

For lumping together two subdomains I will use a negative prefix; thus, non-corrective contrast lumps together oppositive and counterexpectative contrast (Spanish pero).
opposition is obtained both by more extreme position on a scale (*enormous in size* vs. *just tall*) and multiple scales (tallness and arming in 4c vs. only tallness in 4a/b).

(4) Increasingly stronger oppositive contrast

a. *Goliath*<sub>AP</sub> was tall. *David*<sub>OP</sub>, on the contrary, was short.

b. *Goliath*<sub>AP</sub> was enormous in size. *David*<sub>OP</sub>, on the contrary, was a small shepherd boy.

c. *Goliath*<sub>AP</sub> was enormous in size, over nine feet tall, and wore a bronze helmet and armor. *David*<sub>OP</sub>, on the contrary, was a small shepherd boy … was wearing no armor and was armed with nothing but a sling (Pons 2012: 145).

While English uses a general contrast marker *but* (and rarely *on the contrary* in cases of strong oppositive contrast as in (4)), Spanish and German require special corrective contrast markers (*sino* and *sondern*, respectively) and Russian has a counterexpectative contrast marker *no*, as opposed to oppositive and corrective *a*. In Section 4.1 I will argue that the three semantic subdomains in (1)–(3) must be arranged in two dimensions, as shown in Figure 1a (for a similar suggestion, see also Andrason 2020: 21), as opposed to the semantic maps by Malchukov (2004) and Mauri (2008), which arrange them on a line, but in different ways. The relevant sectors of Malchukov’s and Mauri’s semantic maps are displayed in Figure 1b and c with adapted terminology.

Example (5) from Sierra de Juárez Zapotec illustrates how oppositive contrast interacts with selectives and with non-dominant word order. It contains a contrast marker *pero*, borrowed from Spanish, and two occurrences of the selective (“topic marker”, Wälchli 2022; glossed “TOP”) *nna* following each of the contrasted

---

**Figure 1**: Different proposals for semantic maps of contrast.
phrases (underlined). Sierra de Juárez Zapotec has VSO basic word order; that is, the dominant word order is verb-initial, but in (5), the contrasted phrases come first in the sentences. I will argue in this article that oppositional phrases tend to be sentence-initial irrespective of the dominant word order of the sentence. Languages with verb-initial dominant order are most important for demonstrating this, since the main predicate is never part of the oppositional phrase.

(5) Sierra de Juárez Zapotec (zaa-x-bible 40026011)³

\[
\begin{align*}
\text{PL} & \quad \text{enne} & \quad \text{pobre} & \quad nna & \quad \text{tulidàba} & \quad tsè'e & \quad cq & \quad \text{lani} & \quad \text{le}, \\
\text{TOP} & \quad \text{always} & \quad \text{FUT.be.PL} & \quad 3\text{PL} & \quad \text{with} & \quad 2\text{PL} \\
\text{but} & \quad 1\text{SG} & \quad \text{top} & \quad \text{NEG} & \quad \text{always} & \quad \text{PRS-sit-1SG} & \quad \text{with} & \quad 2\text{PL} \\
\end{align*}
\]

‘[[The poor people]AP will always be with you]A, but [[I]OP am not always with you]C.’

Markers such as Sierra de Juárez Zapotec *nna* or Japanese *wa* are often called “topic markers” and they occur both in contrastive and non-contrastive uses (Kuno 1973). Because “topic” is used in many different ways in the literature, the traditional term “topic marker” is confusing. This paper follows Wälchli (2022) in labeling them “selectives”. In a typological investigation of selectives in 81 languages, Wälchli (2022) uses opposition contrast contexts such as (5) to define selectives and investigates to what extent the set of markers thus obtained is also used on subordinate clauses, notably conditional clauses (see also Haiman 1978). Selective constituents (sentence “topics” explicitly marked with selectives) indicate a point of departure (Dooley and Levinsohn 1999: 35) in the sentence from which further common ground can be established, which is why conditional clauses are highly suitable as selective constituents (see Comrie 1986: 86; Lehmann 1974). Wälchli (2022) finds that selectives tend to occur early in the sentence, but in most languages they follow the constituent they scope over, which is the ideal position for avoiding scope ambiguity, given the characteristically high degree of freedom-of-host-selection of selectives (nominal, pro-nominal and clausal constituents). Unlike contrast markers, which usually only occur once in the contrast construction, selectives also occur on the anchor phrase in the anchor sentence (and in many other contexts beyond the contrast domain).

This paper will show that many languages have specific **opposite contrast markers** which have some characteristic properties of selectives, but also that

---

³ 40 stands for Matthew, the 40th book of the Bible, and 026011 for chapter 26 verse 11. The doculect codes by Mayer and Cysouw (2014) are used, which begin with ISO 639-3 codes.
contrast markers and selectives are largely distinct grammatical category types. This result is obtained by means of several intermediate steps. First, the similarity space of contrast markers is explored and oppositive contrast is identified as a fuzzy section of it. Next it will be shown that selectives and non-dominant word order in VSO/VOS-languages are strongly associated with oppositive contrast. In a third step, specific oppositive contrast markers are automatically extracted from parallel texts in a large number of languages. Finally, the properties of these oppositive contrast markers will be compared to those of a set of selectives.

Table 1 summarizes the properties of contrast markers and selectives which are most expectable from earlier literature. The empirical question of the extent to which there are intermediate cases is one of the questions addressed in this paper.

It has been reported for many languages that contrasted phrases tend to be placed sentence-initially. Myhill and Xing (1996) show that oppositive contrast triggers non-dominant object-initial order in Biblical Hebrew (basic VSO/SOV) and Mandarin (basic SVO/SOV). This paper will discuss word order in languages with dominant verb-initial order as in (5), where a difference between dominant and non-dominant order can obtain for all contrasted non-predicative phrases.

Not only the position of the contrasted phrases, but also that of the contrast marker is of interest, as illustrated in (6) from Estonian. Among other things, Estonian aga ‘but’ can occur following the oppositive phrase as in (6a) (glossed BUT2) or, more commonly, at the beginning of the contrast sentence as in (6b) (glossed: BUT1). The position following the oppositive phrase is reminiscent of the typical order of selectives.

**Table 1**: Most expectable properties of contrast markers and selectives.

<table>
<thead>
<tr>
<th></th>
<th>Contrast markers [such as pero in (5)]</th>
<th>Selectives [glossed TOP, such as nna in (5)]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Occurrence</strong></td>
<td>Only once: in the contrast sentence OR at the end of the anchor sentence</td>
<td>In both anchor and contrast sentences</td>
</tr>
<tr>
<td><strong>Word order</strong></td>
<td>Initial in contrast sentence [but also other positions]</td>
<td>Following both anchor phrase and oppositive phrase</td>
</tr>
<tr>
<td><strong>Combinability with contrast markers</strong></td>
<td>No (because it already is a contrast marker by itself)</td>
<td>Yes, often combined with a contrast marker</td>
</tr>
</tbody>
</table>

4 German aber behaves largely the same (see Sæbø 2002).
(6) Estonian (est-x-bible-1997 40026011, 41015023)
   a. [Vaeseid]AP on teie juur-es ju alati,
      poor.PART.PL be.FRS.3SG 2PL-GEN at-INESS PTC always
      [mind]OP aga ei ole te-i-l alati.
      1SG.PART BUT2 NEG be.CONEG 2PL-PL-ADESS always
      ‘... the poor you always have with you, but you do not always have me.’
   b. ... ta-lle pakut-i mürri-veini,
      3SG-ALLAT offer-IMPERSON.PST.SG myrrh-wine.PART
      aga tema ei vôt-nud vastu.
      BUT1 3SG.NOM NEG take-PTCP.PST against
      ‘... they offered him wine mixed with myrrh, but he did not take it.’

If we want to label (6b) as a micro-domain, it might be called “non-responsive contrast”, where the subject of the contrast sentence ‘he[=Jesus]’ is non-obedient or non-cooperative. More important than such a label is that there is a tradeoff between emphasis on the opposite phrase and emphasis on the entire contrast sentence. From this and other examples we can deduce that oppositional and counterexpectative contrast do not strictly exclude each other. Non-responsivity is expressed in the predicate of the contrast sentence, and so it is rather the anchor and contrast sentences as wholes that are opposed to each other. Note also that the sequential character of the two states-of-affairs in (6b) takes away emphasis from the opposition of the referents.

As shown in Table 2, the semantic and pragmatic gradualness of oppositional contrast can be described in terms of degree of emphasis on the contrast between the contrasted phrases. Example (6b) is primarily counterexpectative, but it also holds the possibility of a construal in terms of opposed referents. Put differently, the distinction between oppositional and counterexpectative contrast is fuzzy, which will be reflected by a gradual transition between these two regions of the similarity space.

<table>
<thead>
<tr>
<th>Degree of contrast between contrasted phrases</th>
<th>Strong emphasis (4c), (7)</th>
<th>Emphasis (6a)</th>
<th>Little emphasis (6b) (since no contrast of phrases) (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong oppositional contrast</td>
<td>Merely polar oppositional contrast</td>
<td>Contexts such as non-responsive contrast</td>
<td>No oppositional phrase</td>
</tr>
</tbody>
</table>

Table 2: Semantic/pragmatic gradualness of oppositional contrast between contrasted phrases as degree of emphasis.
Fuzziness goes against Myhill and Xing (1996: 310–311), who suggest a strict definition of oppositive contrast based on the idea that the referents of the contrasted phrases together form a set and that the verbs have opposite or the same meaning, and in the latter case there must be some opposite meaning elsewhere in arguments or adjuncts. Example (6a), as other examples where the predicates in anchor and contrast sentences only differ in polarity, is an instance of merely polar oppositive contrast. Strong oppositive contrast, such as (7) from Catalan (see also (4c) above), provides more prototypical examples of oppositive contrast.

(7) Catalan: strong oppositive contrast (cat-x-bible-inter 42021004)

\[
\text{ella, en canvi, ha donat} \quad \text{el} \quad \text{que} \\
\text{3SG.F in change AUX.PRS.3SG give.PTCP 3SG.M REL} \\
\text{necessitava, tot el que tenia per a viure.} \\
\text{need.PST.IPFV.3SG all 3SG.M REL hold.PST.IPFV for to live.INF} \\
\text{[… all these (=rich people) have given what they had in abundance,] but} \\
\text{she (=this woman) has given what she needed (herself), all what she had for} \\
\text{a living.]}
\]

Example (7) is not about offering versus not offering gifts, but about the least expectable kind of person offering the highest possible amount. Interestingly, this also makes the oppositive referent counterexpectative (as opposed to counterexpectative contrast, where the contrast sentence as a whole is counterexpectative). Also note that (7) is not easily reversible, although interchangeability of order of coordinands is often considered to be a typical property of oppositive contrast (Rudolph 1996: 113).

The rest of the paper is structured as follows. Section 2 provides more background. Section 3 introduces the pipeline of data-and-doculect-sampling. Since the question as to how contrast and selectives relate to each other cannot be addressed in one step, we have to proceed in subsequent steps building on each other in Section 4. Section 5 discusses the results and Section 6 concludes the paper.

2 The syntax and semantics of contrast

2.1 The preselected domain of this study and what is not treated

Contrast entertains many relationships to other domains, not all of which can be treated here. The starting point will be contexts most often encoded by English but across different translations of the gospels, assuring all the three subdomains listed
in (1)–(3) will be included. This choice excludes prototypical instances of concession (*although*), a domain often treated together with contrast (see, for instance, Rudolph 1996). Concession is typically encoded by subordinate clauses whereas contrast constructions are mostly coordinate. However, there are languages, such as Sanuma (Borgman 1990: 59), where contrast is generally expressed by subordinate concessive constructions when marked overtly and this investigation includes translation-equivalents of ‘but’ even if construed as subordination.

Contrast markers frequently grammaticalize from restrictive markers (Malchukov 2004: 194). However, words for ‘only’ are included here only to the extent they happen to be used in oppositional, counterexpectative and corrective contrast.

It is well-known that contrast is closely related to focus, see, e.g., Umbach (2005) — who also emphasizes the close relationship of contrast with denial (see also Spenader and Maier 2009). Dik (1997: 331–335) regards oppositional contrast (“parallel contrast”) and corrective contrast (“replacing contrast”) as types of foci. However, Matić and Wedgwood (2013) show that the notion of focus is problem-ridden from a cross-linguistic perspective both in parametrized approaches (where corrective focus is considered a focus type of its own) and in unified approaches. The consideration of focus is beyond the scope of the present study.

Contrast is also closely related to conjunction (‘and’-coordination; Haspelmath 2004: 5). Russian a, for instance, is often conceived of as intermediate between conjunction (‘and’) and contrast (‘but’). However, as shown by Krejdlin and Padučev (1974), it also has specific syntactic properties in many of its uses, tending to be immediately followed by a topic (“theme”) irrespective of whether the use is contrast or conjunction as in (8), where one constituent from the first conjoined clause is picked up in a new role in the second conjoined clause (see also Jasinskaja and Zeevat 2008).

(8) Russian (rus-x-bible-modern2011 41009007)

\[
\text{Togda javi-l-o-s' oblako i}
\]

\[
\text{then appear-PST-SG.N-REFL cloud(N).NOM.SG and}
\]

\[
\text{nakry-l-o ix svoj-ej ten'-ju,}
\]

\[
\text{cover-PST-SG.N 3PL.GEN POSS-REFL-INS.SG.F shadow(F)-INS.SG a}
\]

\[
\text{iz oblak-a razda-l-sja golos …}
\]

\[
\text{but.OPP out.of cloud-GEN.SG sound-PST.SG.M-REFL voice(M)[NOM.SG]}
\]

‘… there was a cloud that overshadowed them: and a voice came out of the cloud …’

In the literature on information structure, “contrastive topic“, such as that marked by a specific rising intonation contour in English (“[t]he phrase denoting what the question being addressed is about”, Constant 2014: 17), is an important notion.
However, it has to be pointed out that “contrastive topics” also occur in enumeration as in (9) with initial phrases in comparative sequences with more than two coordinands.

(9) Sierra de Juárez Zapotec (zaa-x-bible 66021020)

… ñyyy ñyy nnu cca ñyy nna, lá-q ónice. stone REL be five top is.called-3IMPERS onyx

Nu cca xùppa nna, lá-q cornalina. REL be six top is.called-3IMPERS carnelian

Nu cca gàtsiá nna, lá-q crisólito …

REL be seven top is.called-3IMPERS chrysolite

‘… the fifth sardonyx, the sixth carnelian, the seventh chrysolite …’

The present investigation is restricted to contrast constructions with only two coordinands, which are in-line with Mann and Thompson’s (1988: 248) definition of contrast as always having “exactly two nuclei”.

2.2 Saliency of the oppositive phrase

In Section 1, I have claimed that oppositive contrast, although a meaning, is also associated with a syntactic constituent type: the oppositive phrase. Grammatical markers are not seldom associated with phrase types; case, for instance, usually has a noun phrase host. However, oppositive phrases are special in that (i) they are not definable via dependency on predicates and (ii) there are few restrictions on what kind of constituent can qualify as oppositive phrase. Hence, if the oppositive phrase exists, there are only two alternatives: (a) either it is a pre-established category, such as Rizzi’s (1997) postulated Top(ic)P(hrase) – an option I do not seriously consider – or (b) it is a matter of saliency: it exists if there are properties that render it salient. This section deals with such properties.

A first step is to exclude what cannot be an oppositive phrase. The oppositive phrase is a subpart of the contrast sentence, not the entire contrast sentence or its main predicate or illocution. Put differently, the oppositive phrase cannot contain the main predicate of the sentence or any word directly associated with the main predication such as sentence negation (no verum-elements in the sense of Umbach 2005: 218).

Like conditional clauses, oppositive phrases play a role in stepwise establishing common ground (Lehmann 1974). As such, they will strongly tend to be sentence-initial or immediately follow a sentence-initial contrast marker. This yields Property 1 [P1]: SENTENCE-INITIAL POSITION (Note that a phrase following a contrast
marker also counts as initial). P1 is most salient if the initial position goes against
the expected dominant word order. However, P1 is particularly important in
rendering all non-initial phrases unlikely candidates for oppositive phrases. It
follows from P1 that contrast sentences with initial predication-level elements,
such as finite verbs or sentence negation, are unlikely to contain oppositive
phrases.

Saliency may be strengthened if an initial phrase is detached from the rest of
the sentence by a marker, which yields P2a: OPPOSITIVE-PHRASE-FINAL MARKER, which can be
achieved by a selective as in (5) or an oppositive-phrase-final contrast marker as in
(6a). From P2a, we may derive the hypothesis in (10):

(10) If a language has both contrast-sentence-internal and contrast-sentence-
initial contrast markers, the contrast-sentence-internal one will tend to be
opposite (follow the opposite phrase) and the sentence-initial one
counterexpectative.

This is illustrated in (11) from Somali and (6) from Estonian.

(11) Somali (som-x-bible 40026011, 41015023)
a. … anig-u-se mar wal-ba idin-la-ma joog-o
   1SG-SBJ-butt.OPP time each-also/always 2PL-with-NEG remain.at-NEG.1SG
   ‘[… the poor you always have with you,] but you do not always have me.’
b. laakiin ma uu qaadh-an.
   but.CEXP  NEG 3SG get/pick.up-NEG.PST
   ‘[… they attempted to give him wine mixed with myrrh,] but he did not
take it.’

P2a may suggest that oppositive contrast markers will most likely occur following
oppositive phrases. However, oppositive contrast markers have to single out two
constituents: (i) the oppositive phrase and (ii) the entire contrast sentence. Since
the optimal position for contrast markers is between the anchor and the contrast
sentences, (ii) is in conflict with (i), we may assume that (i) will often prevail over (ii).
However, a sentence-initial contrast marker – even if not fully dedicated to the
expression of oppositive contrast – can single out an oppositive phrase, if the contrast
marker tends to collocate with an immediately following oppositive phrase. This
yields P2b: SPECIFIC CONTRAST MARKER COLLOCATING WITH IMMEDIATELY ADJACENT OPPOSITIVE PHRASE.
If a language has two sentence-initial contrast markers sensitive to the oppositive–
counterexpectative contrast distinction, such as Russian a OPP and no CEXP, it can be
derived from P1 that the former will be immediately followed by an oppositive
phrase more often than the latter. Krejdlin and Padučeva (1974) argue that Russian a
in two of three major uses must be immediately followed by a “theme”, i.e., an oppositive phrase. Examples (12a) and (12b) differ in their construal in two respects that go hand-in-hand: (i) in the use of the contrast marker \( a \) versus \( no \) and (ii) in the presence (underlined) or absence of an initial oppositive phrase. Note that the English translation and (12b) lack oppositive phrases because the contrast marker is followed by predicate-associated elements.

(12) Russian (40007003, 42006041 rus-x-bible-modern2011)

Počemu ty zameča-eš sorink-u
why 2SG notice-PRS.2SG speck-ACC.SG

\( v \) glaz-u (u) (so-)brat-a svo-ego …
in eye-LOC.SG (at) (fellow-)brother-GEN.SG REFL.POSS-GEN.SG

a. … \( a \) u sebja \( v \) glaz-u

\( \text{but.opp} \at\) REFL.GEN in eye-LOC

\( ne \) zamečaes brevn-a?
not notice-PRS.2SG beam-GEN.SG

b. … \( no \) ne zameča-eš brevn-a

\( \text{but.cexp} \) not notice-PRS.2SG beam-GEN.SG

\( v \) svo-ém glaz-u?
in REFL.POSS-LOC.SG eye-LOC.SG

‘And why do you see the speck that is in your brother’s eye, but do not notice the beam of wood that is in your own eye?’

It might be argued that sentence-initial position [P1] affecting the anchor phrase in the anchor sentence may have similar effects indirectly, as in (13). However, if \( me \) in (13) is an oppositive phrase, it is unlikely to interact in long distance with the contrast marker and contrast markers used in such examples are unlikely to be oppositive contrast markers (and, as expected, English \( but \) is a general contrast marker and not an oppositive contrast marker).

(13) English (eng-x-bible-lexham 40026011)

\[ [\text{the poor}]_{ap} \text{you always have with you, but you do not always have me.} \]

Turning to semantics, an oppositive phrase in the contrast sentence is typically opposed to some element in the anchor sentence. Sæbø (2002: 261) calls this element the alternative. Alternatives may be explicit or implicit. However, a referent is at least usually not opposed to itself. Hence, it is reasonable to postulate that the \text{REFERENT OF THE OPPOSITIVE PHRASE CANNOT BE THE ALTERNATIVE} [P3]. In (14) from Ukrainian, \( m_i \) ‘we’ cannot qualify as oppositive referent, because it has the same role in the anchor sentence (same goes for \( you \) in (13)).
In (14), the semantically best candidate for an oppositive phase is the NP ‘this one [=Jesus, this man]’ with the explicit alternative ‘Moses’ in the anchor sentence and, indeed, many translations in many languages have ‘this (man)’ as initial oppositive phrase in the contrast sentence. However, in (14) the initial constituent in the contrast sentence is the entire underlined subordinate clause ‘where this one has come from’, marked with the second position clitic ž with oppositive function (follows the first word of the phrase, glossed as “(2)” in parentheses). It has a partly implicit alternative ‘Moses has come because God sent him to teach us’. Even though we may assume that [P4] EXPLICIT ALTERNATIVES are better than implicit ones, (14) demonstrates that the initial position of the oppositive phrase is more important.

Example (15) from Lithuanian (o is cognate with Russian a) illustrates that what matters is alternative referents and not lexical semantics. The lexeme (aš ‘I’) can be the same if the referents are different (Note P2a indeed in the English translation).

As already pointed out in Section 1, EMPHASIS [P5] is a very important factor. Oppositive phrases are emphatic. Emphasis implies that a meaning is explicitly expressed, which entails that emphasis always has a formal component. Given the same or similar meanings, forms that are more distinctly articulated (longer, prosodically more marked and more clearly detached) are more emphatic. Semantically,
emphasis has various aspects. Particularly important for our purposes are (i) extreme position on quantity scales and unexpectedness/surprise (Trotzke 2017: 35 speaks of “scale of likelihood”).

The list of salient properties of opposite phrases is summarized in Table 3:

Table 3: Properties of salient opposite phrases.

<table>
<thead>
<tr>
<th>Formal properties</th>
<th>Semantic properties</th>
<th>Both semantic and formal</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1: Sentence-initial position, P2a: Oppositive-phrase-final marker, and/or P2b: Specific contrast marker collocating with immediately adjacent oppositive phrase</td>
<td>P3: Referent of the opposite phrase is not the alternative, P4: Preferably has an explicit alternative in the anchor phrase in the anchor sentence</td>
<td>P5: Emphasis</td>
</tr>
</tbody>
</table>

Table 3 demonstrates that formal properties are as important as semantic ones. Oppositive contrast is a matter of family resemblance with more and less prototypical examples. This makes its investigation a challenge, because, as pointed out by Myhill and Xing (1996: 354), there is a risk of circularity. Myhill and Xing’s (1996) approach is to apply a rigid distinction based on two semantic criteria and to not include word order in the definition, which is a feature that their study correlates with oppositive contrast. What is done here instead is to start with the form of contrast markers and to have oppositive contrast emerge as a region of the similarity space of contrast markers. Put differently, I will rely almost entirely on the strength of P2b (and P2a will also play in, because different order of contrast markers, such as Estonian aga1/2, is coded differently). In Section 4.1 we will see that Dimension 2 of the similarity space of contrast markers corresponds to the semantic/pragmatic gradualness of oppositive contrast in Table 2, although all properties in Table 3 except P2a/b are initially ignored.

3 Study design, data and method

3.1 The design of this study and its theoretical and practical implications

The present study is (i) is corpus-based, (ii) considers markers extracted from texts, (iii) makes use of quantitative methods, (iv) uses parallel texts, (v) is massively
cross-linguistic, (vi) uses translations of the New Testament (NT), and (vii) investigates contrast. This section discusses major practical advantages and disadvantages of these choices and their theoretical implications.

(i) The encoding of a functional domain may differ in **degree of language-internal variability** ranging from entirely stable (no variation) to entirely emergent (not conventionalized). Hopper (1998) emphasizes the emergent nature of grammar where structure is always viewed as provisional. In order to make the notion fruitful for typological research it is better to speak of **partially emergent grammar**. There is a constant interplay between discourse and conventionalized structure. Using corpora has the advantage that emergent structure reflected as variation in language use is not precluded from consideration, but has the disadvantage that stable and emergent structure cannot easily be kept apart if only one text per language is available. This could be a major problem, if the aim of the study was to classify languages into types, but no attempt of classification of languages into types is made here. Instead, each text is considered a doculect (a documented language variety) of its own.

(ii) For the purposes of this study, very little annotation and analysis is needed. It is sufficient to identify markers in contrast constructions. No information about how these markers are integrated in language systems is required. The advantage is that even texts in languages with very restricted or no available grammars or dictionaries can be used.

(iii) In linguistics, there has long been a traditional tension between universals and diversity. This study makes use of quantitative methods for data mining that can detect major trends in fuzzy datasets that reflect the full range of attested diversity. I am using such methods because linguistic data are both highly diverse and highly constrained at the same time. The aim is to find strong generalizations in diverse, not strictly implicational, data. A disadvantage is that such methods always come in families of slightly different tools where it is not always easy to find the optimal tool for a certain set of data. Rather than striving at optimizing the choice of specific data mining tools, I will use simple tools yielding sufficiently good results.

(iv) The use of data from parallel texts comes at the cost of translated, rather than original, planned, rather than spontaneous and written, rather than spoken language. The motivation for the choice is that parallel texts allow us to compare languages at the maximally granular level of contextually embedded exemplars rather than on the level of preconceived abstract functional domains. It is often assumed that meanings can be taken for granted as extralinguistic comparative concepts (Haspelmath 2010: 681). However, nothing involved in language is truly extralinguistic. Functional domains relevant for cross-linguistic comparison are not given, but must be established as regions in
similarity space from empirical cross-linguistic investigations. However, functional domains can only be established in cross-linguistic investigations if the unit of comparison is more fine-grained than the functional domain. Since there is no empirical extralinguistic way to describe the degree of similarity between pairs of meanings, similarity space can actually not be fully abstracted from form. This is also why I prefer similarity space over “semantic space” or “semantic map” here.

(v) A massively cross-linguistic approach is adopted, because it is not easily possible to distinguish arbitrary from non-arbitrary structure in specific languages. This study relies on the assumption that cross-linguistically recurrent identity of form reflects similarity in meaning (Haiman 1985: 19). As a consequence, functional domains can be modeled as similarity spaces in which the distance between any pair of aligned exemplars in parallel texts reflects the probability that these are encoded by the same marker in any language (Wälchli and Cysouw 2012). However, it is well-known in typology that cross-linguistically recurrent patterns may also reflect large-scale arbitrary patterns due to language contact or genealogic relationships (see, e.g., Dryer 1992). Considering large and diverse sets of languages minimizes possible hidden effects of large-scale arbitrary patterns. Since it is difficult to remove bias entirely, it can also be useful to compare results in more and less biased datasets as will be shown in Section 4.1.

(vi) The choice of using translations of the New Testament follows from (iv) and (v). It is the only massively parallel text with many electronically available versions even in minor languages from all over the world with sufficiently variable content. One advantage of the New Testament for studying contrast is that there is a number of contexts with strong emphatic opposition of referents, which happen to represent the strongest kind of oppositive contrast (see Section 4.1). However, there is also reason to believe that the use of contrast markers in the Bible translations is not always fully representative of the languages of the translations. Koine Greek de ‘but’ is used extensively in narrative sequence, and translations strongly influenced by Greek or Latin reflect such use, as But when the Pharisees saw it (40012002) in the King James translation. Such contexts are excluded here, as only passages where but occurs in many translations are considered. It also has to be pointed out that the language of the Bible is prestigious in many languages, so that features of Bible translations can spread to other texts. For using Bible translations in typological investigations, see also de Vries (2007).

(vii) Contrast markers are known to be maximally unstable diachronically and are among the “most vulnerable items to contact-related linguistic change in grammar” (Matras 1998: 281). The wide spread of Spanish, Arabic and
Russian contrast markers to unrelated and geographically distant languages makes clear that it is hardly possible in this domain to reconstruct the precolonial situation. However, it is important to notice here that there is not necessarily any correlation between (non-)arbitrariness and stability. In the Bible corpus used here, there is a single one among 27 German translations, the Grünewalder Bible, where aber is largely restricted to second position and used as oppositional contrast marker as opposed to counterexpectative (je)doch ‘but, however’ in initial position. This aber is used as an emergent oppositional contrast marker in a similar way as language-internally much more stable a in East Slavic languages. However, several Polish translations use both ale and lecz in non-oppositional (counterexpectative and corrective) contrast without any clear distinctive pattern. Andrason (2020: 17) shows that lecz is stylistically marked and occurs more often than ale in conservative Bible translations whereas it hardly occurs in informal Polish conversation. Polish lecz can therefore be quite arbitrarily distributed in Polish written texts and its use is not stable. To summarize, contrast is a highly interesting domain to study the interplay between emergent structures and cross-linguistically general trends.

3.2 Data and method

The material used is a corpus of translations of the New Testament originally compiled by Mayer and Cysouw (2014). For a list of 1,659 texts with 1,259 different ISO 639-3 language codes included in this study, see Appendix 3.2.A. However, all but one step are based on much smaller subsets.

This study makes use of several quantitative tools: (i) Principal Coordinate Analysis for constructing and visualizing the similarity space of contrast markers (Section 4.1), (ii) Principal Component Analysis for exploring the different loadings of markers (Sections 4.1–4.2), (iii) Partitioning-Around-Medoids for identifying clusters (Section 4.1) and (iv) t-value as a collocation measure for extracting sets of similar markers in the corpus with a certain distribution profile (Section 4.3).

(i) Principal Coordinate Analysis (PCoA) [R: cmdscale()], also called classic metric Torgenson scaling, is a metric variant of multidimensional scaling (MDS). It is often somewhat sloppily called “multidimensional scaling” in the literature, but differs, for instance, from optimal classification MDS (Croft and Poole 2008; see van der Klis and Tellings 2022 for a survey of linguistic applications of various kinds of MDS). PCoA takes a dissimilarity matrix as input, here calculated with Hamming Distance (see Appendix 3.2.B for further explanation). A major reason
for using PCoA is that it is most similar to Principal Component Analysis among
different kinds of MDS.

(ii) Principal Component Analysis (PCA) [R: prcomp()] differs from PCoA most
markedly by taking a set of variables rather than a distance matrix as input. In
our application, each marker in each language is a variable (see Appendix 3.2.C).
What is called “dimensions” in PCoA is called “components” in PCA. An
advantage of PCA is that the contribution of each variable (i.e., marker) to each
component can be quantified. The technical term for this is “loading”.

(iii) Partitioning-Around-Medoids (PAM) [R: pam()] is a clustering algorithm taking
the same input as PCoA, a dissimilarity matrix. PAM is useful in typology
because it reduces sets of exemplars to clusters roughly corresponding to
functional domains that typologists using description-based comparison tend to
posit as comparative concepts. As we will see, this works very well where
meanings are strictly delineated, but not with fuzzy meanings.

(iv) $t$-Value (Manning and Schütze 1999: 163) is a collocation measure that can be
used to rank the distribution in the parallel text corpus of all salient potential
marker candidates (all wordforms and all character sequences within word-
forms) in all languages according to how well they match a search distribution.
If we can determine a search distribution for oppositional contrast (see Section 4.3
for how this is done), we can search all languages in the whole parallel text
corpus for markers that best match this distribution profile. All markers with
$t$-values above a certain threshold can then be considered good candidates for
expressing oppositional contrast and the automatic output can be verified by
using reference materials. This is an efficient procedure for identifying many
markers belonging to a gram type (a cross-linguistic set of grammatical markers
or “grams” sharing a core of prototypical contexts of use; see Dahl and Wälchli
2016).

Since advanced automatic processing is not possible right from the beginning, this
investigation uses a stepwise procedure, where initial steps taken do not cover
the entire diversity of the corpus, but whose results may then be a basis for later
semi-automatic explorations. This procedure is summarized in Table 4, with steps
feeding into each other consecutively numbered. The choice of number of languages
in each step will be explained in Section 4.

In accordance with the stepwise procedure, Section 4 falls into various
subsections, whose roadmap is anticipated here. For each subsection there are
appendices providing additional data.

In Section 4.1, oppositional and counterexpectative contrast emerge as opposite
poles of Dimension 2 in PCoA analyses of both a European and a world-wide sample.
Europe is chosen first, because this allows for a comparison with Mauri’s (2008, ch. 4) investigation focusing on Europe. It is shown by means of partitioning that the oppositive–counterexpectative dimension does not lend itself easily for data reduction into types, unlike corrective contrast in Dimension 1.

In Section 4.2 it is shown, by means of a principal component analysis, that non-dominant non-verb-initial word order in VSO/VOS languages and selectives pattern together with oppositive contrast.

At this point, oppositive contrast is established as a region of the similarity space of contrast. However, we still have only very few candidates for oppositive contrast markers. These candidates are taken as “seeds” (see Dahl and Wälchli 2016 and, for a similar approach, Asgari and Schütze 2017) for identifying the most prototypical oppositive contrast contexts in the whole NT. With this oppositive contrast prototype domain we can then search the whole corpus for many more oppositive contrast markers. In Section 4.3, the oppositive contrast markers detected are then further coded manually for expectable differences between contrast markers and selectives (Table 1). In Section 4.4, oppositive contrast markers and selectives are arranged in a space defined by these properties, which allows us to determine to what extent oppositive contrast markers differ from selectives.

<table>
<thead>
<tr>
<th>Step</th>
<th>Name</th>
<th>No of NT verses</th>
<th>No of texts</th>
<th>No of lgs</th>
<th>Procedure</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English ‘but’</td>
<td>Whole NT</td>
<td>32</td>
<td>1</td>
<td>Automatic</td>
<td>Frequency</td>
</tr>
<tr>
<td>2</td>
<td>Contrast markers in Europe</td>
<td>101</td>
<td>63</td>
<td>63</td>
<td>Manual</td>
<td>PCoA, PAM, PCA</td>
</tr>
<tr>
<td>3a</td>
<td>Contrast markers world-wide</td>
<td>52</td>
<td>193</td>
<td>193</td>
<td>Manual</td>
<td>PCoA</td>
</tr>
<tr>
<td>3b</td>
<td>Contrast markers without Europe</td>
<td>Same 52</td>
<td>129</td>
<td>129</td>
<td>Manual</td>
<td>PCoA</td>
</tr>
<tr>
<td>4a</td>
<td>Selectives in contrast domain</td>
<td>35</td>
<td>58a</td>
<td>58</td>
<td>Manual</td>
<td>PCA</td>
</tr>
<tr>
<td>4b</td>
<td>Initial OP in languages with dominant verb-initial word order</td>
<td>Same 35</td>
<td>38b</td>
<td>36</td>
<td>Manual</td>
<td>PCA</td>
</tr>
<tr>
<td>4c</td>
<td>Contrast markers, selectives and word order</td>
<td>Same 35</td>
<td>200</td>
<td>198</td>
<td>Manual</td>
<td>PCA</td>
</tr>
<tr>
<td>5a</td>
<td>Seeds for oppositive contrast</td>
<td>Whole NT</td>
<td>50</td>
<td>50</td>
<td>Automatic</td>
<td>Frequency</td>
</tr>
<tr>
<td>5b</td>
<td>Seeds for counterexpectative contrast</td>
<td>Whole NT</td>
<td>50</td>
<td>50</td>
<td>Automatic</td>
<td>Frequency</td>
</tr>
<tr>
<td>6</td>
<td>Extracting oppositive contrast markers</td>
<td>240</td>
<td>1,658</td>
<td>1,258</td>
<td>Automatic</td>
<td>t-value</td>
</tr>
<tr>
<td>7</td>
<td>Properties in oppositive contrast markers</td>
<td>20</td>
<td>249</td>
<td>193</td>
<td>Manual</td>
<td>Four indexes</td>
</tr>
<tr>
<td>8</td>
<td>Properties in selectives</td>
<td>Same 20</td>
<td>63a</td>
<td>63</td>
<td>Manual</td>
<td>Four indexes</td>
</tr>
</tbody>
</table>

*aOnly languages with selectives; bonly languages with dominant verb-initial word order.

Table 4: Pipeline of data-and-doculect-sampling.
4 Results

4.1 The similarity space of contrast markers

4.1.1 The similarity space of contrast markers in Europe

In this subsection, a similarity space is constructed with PCoA from data from 63 European languages (from eight families, 43 Indo-European, see Appendix 4.1.1.A).

There are many contrast contexts in the NT, English *but* ranges from occurrence in 706–2,082 verses depending on translation, and the necessity of manual coding requires sampling. The beginning was entirely Anglo-centric.\(^5\) English was chosen for two reasons. It has a general contrast marker *but*\(^6\) and there are many different translations available, which allows for assessing prototypicality of contexts within a single language before cross-linguistic comparison is started. I first extracted 500 NT verses where *but* is most frequent across 32 English translations (Appendix 4.1.1.B). Next I selected, within those, verses where *but* usually only occurs once, thus – for reasons of convenience in annotation – minimizing the number of verses with multiple contrast contexts. Then I picked the 101 first verses, which were subsequently extracted in the set of 63 European languages that constitute the focus of this section. I later controlled for the effect of the Anglo-centric starting point, and the result is that many corrective contrast contexts would not have made it to the set of most prototypical examples with more diverse sampling (which is not surprising given that corrective contrast is often unmarked cross-linguistically, see Mauri 2008: 127).

Figure 2 displays the two first dimensions of the similarity space of contrast in Europe for six languages (for all languages, see Appendix 4.1.1.D; for other dimensions, see Appendix 4.1.1.E). Every symbol stands for a verse, and same color and shape of symbols indicate that the contrast marker is the same in the docul ect visualized. Color and shape (and legend order) strictly follow frequency in the dataset in order not to suggest semantic patterns by coloring. Latvian has a general contrast marker *bet* ‘but’ as English *but*. On Dimension 1, corrective contrast (negative pole) – illustrated by Romanian *ci*\(^7\) and Estonian *vaid* – is clearly set apart

\(^5\) Anglo-centric starting points are common in typology. Mauri (2008: 309–310) uses an English questionnaire for data collection. Note also that a large proportion of reference grammars are written in English.

\(^6\) Second position on the contrary is extremely rare in the corpus, so it does not matter that it is not included.

\(^7\) Against expectations from Bilbiie and Winterstein (2011), Romanian *iar* is not used for oppositional contrast.
from non-corrective contrast (positive pole). Only a single context in the dataset is intermediate between corrective and non-corrective contrast (*Do not swear falsely, but fulfill your oaths to the Lord*). Koine Greek *alla* encodes corrective contrast, but also covers a part of counterexpectative contrast, and is opposed to *de*.(2). Bulgarian
and Russian both use *a* for oppositive contrast on the negative pole of Dimension 2, opposed to counterexpectative *no* on the positive pole of Dimension 2. However, in Russian corrective contrast goes with *a*; in Bulgarian it is split.

A major result of the PCoA analysis is that Dimension 2 largely corresponds to the semantic/pragmatic gradualness of oppositive contrast as postulated in Table 2, which is shown schematically in Table 5 (see Appendix 4.1.1.C for more details).

For assessing which markers express oppositive contrast we can use Principal Component Analysis. Principal Component Analysis [PCA] (prcomp() in R) provides us largely with the same result as PCoA (but “dimensions” are called “components” in PCA). However, PCA also assigns all variables (here contrast markers) negative or positive loading values for each component, which allow us to assess how strongly each contrast marker correlates with the component (see Appendix 4.1.1.G for the entire list). Table 6 lists selected doculects which all have oppositive contrast markers marked in boldface (they are all associated with the negative pole of Dimension/Component 2 when loadings in the Principal Component Analysis are considered).

Partitioning-Around-Medoids [R-function pam()] is the statistical equivalent to the common typological practice of assigning markers to rigid functional domains. Figure 3 bottom right shows how the R-function pam() clusters the distance matrix with three clusters.² Three clusters are chosen for comparison with Mauri (2008), who uses the three comparative concepts oppositive, counterexpectative and corrective contrast. The three clusters obtained with Partitioning-Around-Medoids reflect the most optimal classification of contexts by means of rigid functional

| Table 5: Semantic/pragmatic gradualness of oppositive contrast in Dimension 2. |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| **Strong oppositive contrast >** | **Merely polar oppositive contrast >** | **Contexts such as non-responsive contrast >** | **No or dummy oppositive phrase** |
| Example (shortened if given earlier) | (7) rich people give of abundance, this poor woman put in everything | (6a) poor always, me not always | (6b) they offer wine, he not take | … this must happen, but is not yet the end (40024006) |
| Value on Dim2 | –0.301 | –0.162 | 0.175 | 0.237 |
| Catalan (see Figure 3) | *en canvi2* | *en canvi1* | però | |
| Russian | *a* | | | |
| Koine Greek | | | | *

² Eyeballing wrongly suggests a rupture around –0.1 on Dimension 2. No doculect makes exactly that distinction (see Appendix 4.1.1.D) and not even the distinction between Clusters 1 and 3 in partitioning reflects it strictly.
subdomains for oppositive, counterexpectative and corrective contrast. Blue color in Table 6 indicates all fields where an oppositive contrast marker (boldface) holds the majority within a subdomain. If oppositive contrast markers were equally distributed across doculects, blue color would be equally distributed across all rows of Table 6, which is not the case. Rather, oppositive contrast markers range from markers such as Koine Greek *de(2)* and Slovenian *pa2* (top rows in Table 6), which are most frequent even in the counterexpectative cluster, to doculects such as Basque, where oppositive contrast markers are not even dominant in the oppositive contrast cluster.9

9 Not unexpectedly, Mauri (2008) does not treat oppositive contrast markers consistently. For instance, Albanian *kurse* is assigned to oppositive contrast, but Basque *ordea2* and *berriz2* or Catalan *en canvi1/2* are not detected.

Figure 3: Similarity space of the contrast domain in European languages (further doculects).
We may conclude that Dimension 2 is gradual in the sense that markers that support Dimension 2 range from rare oppositive contrast markers (mostly restricted to strong oppositive contrast) such as Catalan en canvi to frequent markers such as Slovenian pa2. The large difference between Basque ordea2 and berriz2 and Catalan en canvi/2, on the one hand, and Slovenian pa2, on the other hand, can also be seen in Figure 3 on Dimension 2 (y-axis) of the PCoA-plots for Basque, Catalan and Slovenian.

While corrective versus non-corrective contrast appears as a rather strict distinction in the similarity space, it is not entirely strict. There are minor expressions for ‘on the opposite’, as, for instance, Russian naprotiv [one occurrence]. But more importantly, there are more intermediate contexts than the PCoA plots suggest. They did not make it into the sample, because they happen to be non-prototypical ‘but’-contexts. Put differently, not only can they be encoded both by typical contrast and non-contrast markers, but they are often left unmarked or expressed by markers meaning ‘only’, ‘instead’ or ‘rather’, such as Spanish No teman a los que … Teman solo a Dios … (40010028) ‘And do not be afraid of those [who kill the body but are not able to kill the soul,] but instead be afraid of the one [who is able to destroy both soul and body in hell.]’ These examples often encode partial correction (hence ‘only’).

The difference in gradualness between the two different dimensions is of uttermost methodological importance. Corrective markers tend to be strictly delimited in prototypical contrast contexts, whereas the reason why they are detectable by means of more
general comparative concepts, such as those used by Mauri (2008), whereas this is considerably more difficult for the fuzzier oppositive contrast markers. This is illustrated in Figures 4 and 5 where contrast markers with high opposite loadings are shown. The x-axis shows contexts in ranking order according to the single relevant dimension (Dimension 1 for corrective contrast and Dimension 2 for oppositive contrast). Figures 4 and 5 only display languages that reflect the relevant distinction.

**Figure 4:** Corrective versus non-corrective contrast in Dimension 1 in selected European doculects.

**Figure 5:** Contrast markers sensitive to Dimension 2 in selected European doculects.
Figure 4 shows that corrective (red) versus non-corrective contrast (black) is a rather strict distinction and Figure 5, where various kinds of oppositive contrast markers are plotted in different colors, shows that oppositive contrast markers are distributed in a much fuzzier manner.

4.1.2 Expanding to a world-wide sample

A set of 101 situations could not be handled practically with manual annotation in a larger world-wide sample of 193 languages (see Appendix 4.1.2.A) and was reduced to 52 situations (see Appendix 4.1.2.B). The sample is a diverse convenience sample and is biased towards languages with selectives and languages with verb-initial dominant word order, which will be needed in Section 4.2.

The first two dimensions of the similarity space are largely the same as for Europe irrespective of whether it is constructed with all 193 languages or only with the 129 non-European ones, which is shown in Figure 6 top left where arrows link the two similarity spaces built by PCoA. Each arrow stands for a contextually embedded situation and the arrows go from the space built without European data to the space including European data. Different colors help distinguish the arrows from each other. The two spaces are not completely identical, but very similar as shown by a congruence coefficient of 0.99 (see Mair et al. 2022: 39). Bulgarian (top right) is given as an example from the European dataset. This and further maps in this section are based on the 193-language sample.

Let us now consider selected languages illustrating an oppositive–counter- expectative contrast distinction on Dimension 2 (for all languages, see Appendix 4.1.2.C; for evaluation of the dimensions, Appendix 4.1.2.D).

The following two examples demonstrate how PCoA-plots can contribute to description in particular languages.

Somali (bottom left) – the discussion is picked up here from Section 2.2 (example (10)) – has sentence-initial laakiin ‘but’, borrowed from Arabic, and second-position (Saeed 1999: 92) -se ‘but’ (rarely combined as laakiinse ‘but[... on the other hand]’). Mauri (2008: 152) classifies both markers as counterexpectative contrast, but -se is rather oppositive, which can be clearly seen in the similarity space.

---

10 The order of poles of the two dimensions is adjusted by flipping positive and negative values if necessary; poles are not per se meaningful and can easily switch in PCoA. Tuvinian is excluded from the non-European sample due to borrowing of Russian marker.

11 Since rotation is negligible, this comes very close to a Procrustes analysis (Mair et al. 2022: 37) that would provide a very similar plot.
In Cherokee (bottom right), =skinii ‘but’ (Montgomery-Anderson 2008: 158) is used for oppositive and corrective contrast. The clause-initial conjunctive adverb aseèhno ‘but/however’ links clauses (Montgomery-Anderson 2008: 535) and can be identified as counterexpectative.

Let us now turn to the discussion of a few languages where different markers interact with different kinds of word order.

Malagasy (Figure 7 top left) is particularly interesting because of its basic VOS order and because it has been argued to have clause-final topics (Pearson 2005). However, oppositive phrases are sentence-initial, combined with postposed kosa2 ‘on the contrary’ (Rahajarizafy 1960: 116) – illustrated in (16) – fa … kosa2 (both clearly associated with the oppositive contrast pole in Figure 7) or with bare fa with an initial NP. Bare fa ‘but; because; COMPLEMENTIZER’ is not dedicated to any specific subdomain within contrast (since different word orders with fa are not indicated in

**Figure 6:** Similarity space of contrast markers based on a 193-language world-wide sample.
Figure 7: Further evidence for oppositive contrast markers from the world-wide sample.

Figure 7), whereas *(ka)nefa ‘however’ and *saingy ‘but aside from that’ are clearly counterexpectative.\(^{12}\)

(16) Plateau Malagasy (plt-x-bible-interconfessional 42007046)

\[
\begin{align*}
\text{Ianao} & \quad \text{tsy} & \quad \text{n-anosotra} & \quad \text{diloilo} & \quad \text{ny} & \quad \text{loha-ko;} \\
2\text{sg} & \quad \text{not} & \quad \text{PST-anoint} & \quad \text{oil} & \quad \text{DEF} & \quad \text{head-poss.1sg}
\end{align*}
\]

\[
\begin{align*}
\text{izy} & \quad \text{kosa} & \quad \text{n-anosotra} & \quad \text{ilomanitra} & \quad \text{ny} & \quad \text{tongotro} \\
3\text{sg} & \quad \text{but.opp} & \quad \text{PST-anoint} & \quad \text{perfume} & \quad \text{DEF} & \quad \text{foot-poss.1sg}
\end{align*}
\]

‘You did not anoint my head with olive oil, but she anointed my feet with perfumed oil.’

---

12 If the oppositive phrase is initial, the same often holds true also of the anchor phrase, which is the case in (16) *[ianaol]. This phenomenon is so common that it will not be explicitly discussed for all examples where it occurs.
Maori (Figure 7 top right) has a rich set of markers (Mauri 2008: 152 only lists engari). Of particular interest is Maori ko (called “topicalizing particle” in Bauer 1993: xvii, 398), which triggers subject-initial word order, so called “ko-fronting” (basic word order is VSO). However, atypically for a topic marker, it can mark contrast on its own without another contrast marker present. As can be seen in (17), ko may occur both in anchor sentences and contrast sentences (for further functions, see Pearce 1999). Maori ko has some properties of a topic marker and some properties of a contrast marker, without being a prototypical instance of any of them (and, as can be seen in Figure 7, it is not entirely restricted to oppositional contrast).

(17) Maori (mri-x-bible 40024035)

\[
\begin{align*}
\text{Ko} &\quad \text{te} &\quad \text{rang}i &\quad \text{me} &\quad \text{te} &\quad \text{whenua} &\quad \text{e} &\quad \text{pahemo}, \\
\text{“top”} &\quad \text{DEF} &\quad \text{sky} &\quad \text{with} &\quad \text{DEF} &\quad \text{earth} &\quad \text{TMA} &\quad \text{pass.away} \\
\text{ko} &\quad \text{aku} &\quad \text{kupu} &\quad \text{ia} &\quad \text{e} &\quad \text{kore} &\quad \text{e} &\quad \text{pahemo}. \\
\text{“top”} &\quad \text{1SG} &\quad \text{word} &\quad \text{3SG} &\quad \text{TMA} &\quad \text{NEG} &\quad \text{TMA} &\quad \text{pass.away}
\end{align*}
\]

‘Heaven and earth will pass away, but my words will never pass away.’

Ama is particularly interesting in two respects. The contrast clitic -su following NPs and ulai ‘but (signaling the unexpected)’ (Årsjö 1999: 84, 33) are oppositional and counterexpectational only as a tendency (see Figure 7 bottom left) and Ama can combine -su with a following selective mo TOP, as shown in (18). As in other languages where such a combination occurs, selectives tend to follow oppositional-phrase-final oppositional contrast markers.

(18) Ama (amm-x-bible 40024035)

\[
\begin{align*}
\text{Yo-su} &\quad \text{mo} &\quad \text{nuwoi} &\quad \text{muwoi}. \\
\text{1SG} &\quad \text{but.TOP} &\quad \text{TOP} &\quad \text{certain.day} &\quad \text{NEG}
\end{align*}
\]

‘But I have no certain time.’

Tzotzil illustrates the complex interplay of different contrast markers with different word orders and a marker with properties of a selective. Tzotzil has VOS basic order, but contrasted phrases in oppositional contrast are placed before the verb, usually followed by a final particle =e (Aissen 1992: 49; glossed FIN). The final particle behaves like a selective except that it also often occurs sentence-finally.\(^{13}\) The most frequent contrast marker is pero (from Spanish), used both with and without initial NPs. Since word order is not shown in Figure 7, pero extends over both oppositional and counterexpectational contrast. However, there is also the oppositional contrast marker yan ‘but’ illustrated in (19a). Example (19b) is a non-

\(^{13}\) The initial topic marker a (Aissen 1992: 49) does not occur in the Tzotzil NT.
responsive example with *pero*, which, however, has an initial oppositive phrase marked with \(-e\):

(19) Tzotzil (tzo-x-bible-chamula 40026011, 41015023)

a. 

\[ Ti \quad \text{buch’u-tic} \quad \text{abol} \quad s-ba-iqu=e \quad \text{li’} \]
\[ \text{DEF} \quad \text{REL-PL} \quad \text{poor} \quad 3\text{-REFL-PL}=\text{FIN} \quad \text{here} \]
\[ \text{ono’ox} \quad \text{oy} \quad \text{ta} \quad \text{a-tojol-iqu}=e. \]
\[ \text{always} \quad \text{EX} \quad \text{OBL} \quad \text{2-in.presence-PL}=\text{FIN} \]
\[ \text{Yan} \quad \text{ti} \quad \text{vu’un}=e \quad \text{mu} \quad \text{scotol-uc} \quad \text{c’ac’al} \]
\[ \text{but.opp} \quad \text{DEF} \quad 1\text{SG}=\text{FIN} \quad \text{NEG} \quad \text{all-every} \quad \text{day} \]
\[ \text{li’} \quad \text{oy-un} \quad \text{ta} \quad \text{a-tojol-iqu}=e. \]
\[ \text{here} \quad \text{EX-PTC} \quad \text{OBL} \quad \text{2-in.presence-PL}=\text{FIN} \]

‘… the poor you always have with you, but you do not always have me.’

b. 

\[ Tey \quad \text{laj} \quad y-ac’-be-ic \quad y-uch’ \quad \text{ti} \quad \text{Jesus}=e \]
\[ \text{there} \quad \text{AUX.} \quad \text{fin} \quad \text{PST.3-give-IO-PL} \quad \text{PST.3.SG-drink} \quad \text{DEF} \quad \text{Jesus}=\text{FIN} \]
\[ \text{jun} \quad \text{ch’ail} \quad \text{pox} \quad \text{ti} \quad \text{cap-al} \quad \text{ta} \quad \text{mirra}=e. \]
\[ \text{one} \quad \text{liquor} \quad \text{DEF} \quad \text{mix-POSS} \quad \text{of} \quad \text{myrrh}=\text{FIN} \]
\[ \text{but} \quad \text{DEF} \quad \text{Jesus}=\text{FIN} \quad \text{NEG.EX} \quad \text{AUX.} \quad \text{fin} \quad \text{PST.3.SG-drink} \]
\[ \text{Pero} \quad \text{ti} \quad \text{Jesus}=e \quad \text{mu’yuc} \quad \text{la} \quad \text{y-uch’}. \]

‘… they attempted to give Jesus wine mixed with myrrh, but Jesus did not take it.’

In this section we found that the same two dimensions that characterize the contrast domain in Europe are equally manifest in a world-wide sample. As in Europe, the distinction between corrective and non-corrective contrast is more rigid (Dimension 1), and the oppositive–counterexpectative distinction is fuzzy (Dimension 2). This section has also illustrated a number of languages where contrast markers interact with different word orders and with selectives or selective-like markers. Such languages can also have dedicated oppositive contrast markers, such as Malagasy *kosa*2 and Tzotzil *yan*.

### 4.2 The oppositive-versus-counterexpectative contrast dimension in relation to selectives and sentence-initial oppositive phrases

This section explores the relationship of the oppositive–counterexpectative contrast dimension with word order and selectives in the world-wide sample introduced in Section 4.1.2. For this purpose we will zoom in on non-corrective contrast, which means that all examples with negative values on Dimension 1 in Section 4.1 are
As we have seen in Section 4.1, corrective contrast is a sharply delimited cluster whose examples are not particularly sensitive to the oppositive–counterexpectative distinction. This leaves us with 35 examples per doculect, which are now also coded for presence or absence of selectives in a set of languages with selectives (Appendix 4.2.A) and non-predicative initial word order in VSO/VOS-languages (Appendix 4.2.B).

Example (20) illustrates presence and absence of a selective (TOP) from Hills Karbi.

(20) Hills Karbi (mjw-x-bible 40026011, 41015023)

a. Athë-ke ke-duk kelak atum ke kaike-ta
   thus-TOP NMLZ-poor weary PL top always-ADD
   nang-tum a-long do-ver-ji,
   2-PL POSS-LOC be-always-IRR
   bonta ne-ke nang-tum a-long kaike do-ver-ve.
   but 1SG-TOP 2-PL POSS-LOC always be-always~NEG

   ‘… the poor you always have with you, but you do not always have me.’

b. … bonta Jisu la jun-je-det-lo.
   but Jesus this drink~NEG-PFV-REAL
   ‘[they attempted to give him wine mixed with myrrh,] but he did not take it.’

In non-corrective contrast, the oppositive–counterexpectative distinction is the dominant signal. Hence, the first component of a Principal Component Analysis will reflect it and it will also show how selectives and word order relate to it when this data is added as additional variables.

Appendix 4.2.C lists all PCA-loadings for Component 1, and it turns out that all but one selective [TOP = “topic marker”] values and all non-dominant non-verb-initial word-order-values are negative in the same way as oppositive contrast markers, such as Russian a, which means that the hypothesis of a correlation between oppositive contrast, selectives and initial oppositive phrases is fully confirmed. The loadings of all markers are plotted on the y-axis in Figure 8. Markers and constructions associated with oppositive contrast (selectives, non-initial phrases in dominant verb-initial word order and oppositive contrast connectives) – all grouped to a block in the legend – appear in the lower part of the figure as opposed to counterexpectative contrast connectives in the upper part.

There are also some differences between oppositive contrast markers, selectives and non-dominant initial word order. Clausal oppositive phrases are more prone to

---

14 Generally, selectives and exceptional non-verb-initial word order are not particularly common in corrective contrast.
be initial than to host oppositive contrast markers. This is expected, since initial adverbial clauses are very common even in verb-initial languages. There are many infrequent oppositive contrast markers (few tokens, to the left in Figure 8), whereas selectives and initial oppositive phrases tend to be more frequent across the contrast domain (to the right in Figure 8). In languages with two types of marking strategies, the occurrence of an oppositive contrast marker is almost always combined with an initial oppositive phrase in VSO/VOS-languages, see, for instance, Tzotzil (19).

The only doculect in the sample lacking non-dominant non-verb-initial word order in oppositive contrast is Contemporary Welsh. The Welsh translation from 2013 has only one example with an initial oppositive phrase among contexts sampled, and this example has an initial adverbial subordinate clause, where initial order is expected. However, the translation from 1804 (a modified version the Morgan-translation from 1588, which is still close to Middle Welsh) has many cases of initial oppositive phrases, with the translation from 2004, with some conservative elements, being intermediate. The Irish, Waray and Car Nicobarese translations have low incidences of initial oppositive phrases. Car Nicobarese pöri ‘but’ (21a) usually follows the predicate (Braine 1970: 227). However, in two of the most prototypical

![PCA loadings (y-axis)](image)

**Figure 8:** Markers and constructions sensitive to the oppositive–counterexpectative distinction.

---

15 Welsh has been subject to complex word order change (see, e.g., Willis 1998). From Old Welsh VSO, Middle Welsh developed dominant V2-order in affirmative main clauses from cleft-constructions. After the loss of preverbal particles and as preverbal subject clitics were reanalyzed as the affirmative particles *mi* and *fe* in the 18th century, Contemporary Modern Welsh ended up with rather strict verb-initial word order (and the literary “abnormal” V2-order is associated with English influence).
oppositional contrast examples (see 21b), it follows an oppositive phrase rather than the predicate.

(21) Car Nicobarese (caq-x-bible 42007046, 42021004)

a. Röö yih meen hacham elkui Chu tō tavti;
   not come 2SG anoint head 1SG.OBL AGT/INS coconut.oil
   hachämö kun-röön póri Chuö tō ngam
   be.anointed little-foot(=toe) but 1SG.SBJ AGT/INS that.SG
   tötahūsa tavti.
sweet.smelling coconut.oil
   ‘You did not anoint my head with olive oil, but she anointed my feet with perfumed oil.’

b. … ṛōkhōre chaa nō i kanōlō-re nō
   all 3PL PTC PREP wealth-REFL PTC
chap-hōt ka-hētō-re;
pick.up-COVERED PREFIX-bring-REFL
ngōh kikānō póri nō keu-heūt-hōt-re tt,
this woman but PTC take-finished/all-COVERED-REFL hand/PTC
tō nup tōng unhọ̄īmō ʻō.
AGT/INS DEM merely life/breath 3SG.OBL
   ‘… these all put gifts into the offering out of their abundance, but this woman out of her poverty put in all the means of subsistence that she had.’

No systematic coding of initial oppositive phrases in OVS-languages was made, but “fronting” in contrast is common in Hixkaryana (OVS; Derbyshire 1979: 72), illustrated in (22). The adversative mak ‘contrary to expectation’ is placed after or before the verb sentence-medially, but oppositive contrast also tends to have an initial phrase followed by particles.

(22) Hixkaryana (hix-x-bible.txt 43012008)

Uro haxa ryhe, tano roro-hra mak wehxaha
   I but.OPP EMPH here all-NEG but I.am
   ‘But I am not always here.’

Hixkaryana maintains a flavor of OVS by often using initial cataphoric pronouns with the co-referential lexical item placed finally: Moson haxa ryhe […] wosā ‘this.one but.OPP EMPHASIS … woman’.

In word order typology, a distinction is sometimes made between languages with given–new order (as in Slavic) and languages placing “indefinite, new, or less expected information first”, such as (23) Tohono O’odham (Payne 1987: 802). Interestingly, this does not result in different word order preferences in oppositive contrast.
Both kinds of languages have initial oppositive phrases, which is fully compatible with the function to “instruct the hearer to mentally ‘tag’ the entity as something to be available for deployment” (Payne 1987: 798–799); this is tantamount to saying that common ground is established stepwise.

4.3 Extracting oppositive contrast markers with high cue validity

This section sets up an automatic search procedure for salient oppositive contrast markers across all doculects of the Bible corpus. Methodologically, this implies shifting from manual annotation to methods relying on automated discovery, which allow for larger samples of doculects. The procedure (inspired by Dahl and Wälchli 2016 and Wälchli 2019) consists of the following steps:

(a) Collect “seed” markers for oppositive contrast (markers for which there is previous evidence that their distribution singles out oppositive contrast) from the items with top-ranked loadings in the Principal Component Analysis in Section 4.2 (Appendix 4.3.A).

(b) Do the same thing for counterexpectative contrast, because we need a filter that removes all markers predominantly used in the counterexpectative contrast domain (Appendix 4.3.B).

(c) Compile two sets of verses with most “seed” markers for (a) and (b), which define in which contexts oppositive contrast markers are expected to occur and not to occur (Appendix 4.3.C/D).

(d) Extract all markers that collocate with the oppositive contrast domain above a certain threshold16 and that collocate more strongly with it than with the counterexpectative contrast domain, and let the counterexpectative contrast domain be stronger (i.e., larger) for eliminating any risk that markers not

16 The same collocation measure, t-value, is used both for both domains and the threshold is set at $t = 3.5$. 
primarily restricted to oppositive contrast will be extracted by coincidence (implemented in a Python-program).

(e) Check manually with available reference grammars and dictionaries whether the extracted markers indeed reflect oppositive contrast markers.

The search procedure is designed to have high accuracy (general and counter-expectative contrast markers are reliably removed), but limited coverage: only salient markers with high cue validity make it to the sample. Notably, affixes will be extracted to a much lesser extent than wordforms. Since the purpose of this study is to establish oppositive contrast markers as a category type rather than detecting all languages where they occur, it is sufficient to find many.

The procedure does not contain any filter to exclude selectives, which is why selective-like markers may occasionally be extracted. However, most selectives will not reach the collocation threshold, since they are not sufficiently dedicated to contrast (nearly all selectives are very frequent).

Only 36 of the 50 seed markers in (a) make it to the extraction (Appendix 4.3.F), but it is important to note that extracted markers can cover very different zones of Dimension 2 in Section 4.1, ranging from markers largely restricted to strong oppositive contrast, such as Basque berriz, to markers covering a broad zone such as Koine Greek de(2) (see Table 6). Collocation measuring applies on the level of verses and no filter prevents them from occurring in the anchor sentence (but most extracted markers are in practice restricted to contrast sentences).

Markers in 255 doculects from 199 different languages were extracted. In several languages, extraction is not constant across doculects, which testifies to the large amount of language-internal variability of oppositive contrast markers. For instance, French tandis is extracted only in one of 17 translations and Italian invece only in two of seven translations (see Appendix 4.3.G).

Extraction of counterexpectative contrast markers (checked also with reference grammars and dictionaries, wherever available) reveals that in all doculects with extracted oppositive contrast markers another or several other markers are used in counterexpectative contrast contexts (see Appendix 4.3.F). There is one minor type of wrongly extracted markers: emphatic personal pronouns of the third person singular, such as Xhosa (xho) yena, in only six languages (see Appendix 4.3.E). Emphatic personal pronoun errors are not unexpected since emphasis of oppositive phrases clearly goes hand-in-hand with oppositive contrast.

For further exploration, 20 verses where both oppositive phrases and anchor phrases can be quite easily identified (see Appendix 4.3.H) have been manually coded across all doculects with extracted markers for four properties ((b)–(d) are taken from Table 1). On the basis of the 20 occurrences in the 20 selected verses, the four following indexes are calculated (divided by the number of attested occurrences,
hence ranging between 1.0 and 0.0) Figure 9 displays all four properties and shows that no property always assumes extreme values:

(a) **Frequency**: whether the marker is found in that verse [0–20 tokens indicated by size of circle in Figure 9]

(b) **Word order**: whether the marker occurs before (red in Figure 9, e.g., Russian _a_) or after the opposite phrase (blue in Figure 9, e.g., Georgian _-ki_; phrase-internal order as in Koine Greek _de(2)_ is treated the same way as OP-final, since the two cannot be distinguished in phrases consisting only of one word);

(c) **Combinability with contrast markers**: whether it combines with (left on x-axis in Figure 9) or does not combine with another contrast marker (right on x-axis in Figure 9),

(d) **Occurrence**: whether it behaves selective-like in also occurring with the anchor phrase in the anchor sentence (bottom on y-axis in Figure 9) or not (top on y-axis in Figure 9). Examples with marking only on the anchor phrase, but not on the opposing phrase are also counted as selective-like.

All values are also given in Appendix 4.3.1.

The background of Figure 9 is a two-dimensional logarithmic histogram or heatmap where darkness on a grayscale reflects density of markers. The black square in the upper right corner shows that this area hosts the majority of extracted markers (almost all red, i.e., OP-initial). Table 7 shows that the dominant values are (i) not combined with other contrast marker (right), (ii) no marking on the anchor phrase (top) and (iii) initial position (red). However, if the construction contains another contrast marker (left), the opposite contrast marker strongly tends to follow the opposite phrase (blue dominates in Figure 9 except in the top right corner).

The Batakic markers Batak Dairi (btd) _ukum_ and Toba Batak/Batak Angkola (bbc/akk) _anggo_ stick out (red top left in Figure 9). They are OP-initial (red), mostly combine with another contrast marker and sometimes also occur before anchor phrases. The Batakic markers are also the only ones deviating from (24):

(24) **Universal Trend I**: Phrases with opposite contrast markers almost always occur initially in the contrast sentence.

(24) is dominant even in Batakic (see also van der Tuuk 1864: 364), but, as illustrated in (25), an opposite phrase can also, more rarely, occur sentence-finally in some kind of antitopic-position. This is a rare case where [P1] Sentence-initial position in 2.2 is violated.

(25) **Toba Batak** (bbc-x-bible 42005033)

```
\[ … alai mangan minum do anggo sisean-mi \]
\[ but eat drink AFF but.opp/as.for/if disciple-Poss-2sg \]
\[ ‘but your disciples are eating and drinking’ \]
```
Properties of extracted oppositive contrast markers.

**Table 7:** Distribution of properties of extracted oppositive contrast markers across 20 examples in all doculects.

<table>
<thead>
<tr>
<th>Complex construction with another contrast marker</th>
<th>Simple construction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Postposed</td>
</tr>
<tr>
<td>Only with OP</td>
<td>386 [11.1%]</td>
</tr>
<tr>
<td>(27 lgs)</td>
<td>10 lgs)</td>
</tr>
<tr>
<td>Both AP+OP</td>
<td>116 [3.3%]</td>
</tr>
<tr>
<td>(also includes AP)</td>
<td>8 lgs)</td>
</tr>
</tbody>
</table>

Verses in all doculects added up (maximally 20 per doculect). Dominant values per language in parentheses. A language is counted several times if different doculects have different dominant values.

If the extracted oppositive contrast marker and another contrast marker in the same language occur in different positions relative to the oppositive phrase, the oppositive contrast marker is usually OP-final or OP-internal (26):

(26) Universal Trend II: oppositive contrast markers are more likely to have OP-final order than other contrast markers
restates the hypothesis formulated in (9) in Section 2.2. An exception is the Western Arrarnta initial oppositive contrast marker *kanha* whereas another more general contrast marker *pula* occurs in second position.

In the 20 oppositive contrast contexts surveyed, oppositive phrases are almost always the initial phrase in contrast sentences if combined with one of the extracted oppositive contrast markers. This does not entail, however, that all oppositive contrast markers are immediately followed by an oppositive phrase if these markers are used in entirely different functions.

Universal Trend III: Oppositive contrast markers are almost always immediately adjacent to oppositive phrases in prototypical oppositive contrast contexts

Very few examples go against the trend in (27). One is Bora (28), where *áánetu* (Thiesen and Weber 2012: 290) does not always have an immediately following oppositive phrase. The word order is similar to that of English … but you do not always have *me*, but English is no exception to (27), because *but* is no oppositive contrast marker. All exceptions, such as Bora (28), have initial contrast markers.

(28) Bora (boa-x-bible 43012008)

… *Áánetu* *tsä* *pa-ǐyú-vá-rê* *ámúha-ma* *q* *įjcyá-i-tyú-ne.*

*but* opp NEG all-day-ref-only 2pl-with 1sg.sbj be-fut-NEG-sg.m

‘… but you do not always have *me*.’

To summarize, many languages have salient oppositive contrast markers. The simple automatic search algorithm used here identifies them in more than a seventh of the languages of the corpus in languages from all continents.

4.4 The relationship between oppositive contrast markers and selectives

According to Wälchli (2022), contrast markers only occur once in the contrast construction (within or immediately preceding the contrast sentence), whereas selectives also occur in the anchor sentence. However, some oppositive contrast markers extracted in Section 4.3 sometimes occur even with anchor phrases in the anchor sentence, which suggests that there might be a continuum between oppositive contrast markers and selectives. Let us now look at how a set of 63 selectives (Appendix 4.4.A) behaves under exactly the same conditions. Figure 10 displays the same properties as Figure 9. Its left hand side maps selectives only, whereas its right hand side displays both oppositive contrast markers and selectives. Symbols for selectives are distinguished by light blue outline. The comparison reveals that
selectives tend to occur following anchor phrases in anchor sentences in at least one of three cases (index value on y-axis below 0.66), in word order mostly follow the contrasted phrase (blue) and mostly occur in combination with a contrast marker (left).

While Figures 9 and 10 show that opposite contrast markers and selectives generally tend to have different properties, there are a small number of extracted opposite contrast markers that actually have properties similar to selectives.

There is only one selective also extracted as opposite contrast marker in Section 4.3: Hills Karbi (mjw) ke. However, the extraction will have missed some intermediate markers (for instance, Maori ko, see (17)).

To summarize, although opposite contrast constructions host both opposite contrast markers and selectives, these two types of markers do not have the same distribution within opposite contrast constructions. Whereas opposite contrast markers tend to be strongly restricted to occurring with opposite phrases in contrast sentences, selectives also occur with anchor phrases in anchor sentences. However, since markers differ in how often they occur with anchor phrases, there is a continuum between opposite contrast markers and selectives, but the intermediate area is not particularly crowded. In most cases, opposite contrast markers and selectives can be clearly distinguished.

**Figure 10:** Ratio of occurrence with AP: selectives (left) and opposite contrast markers with selectives (right).

17 Sochiapam Chinantec (Tcso) ne³ is an outlier.
5 Discussion

5.1 The interplay of oppositive contrast markers, selectives and word order

In Section 4.2, it was shown that oppositive contrast markers, selectives and initial non-predicative constituents in VSO/VOS-languages all occur in the oppositive contrast domain. This overlap in use is evidence for some sort of interplay; it clearly demonstrates that the three phenomena types overlap in use. However, the categories neither strictly condition each other nor are they all conditioned by a single underlying syntactic structure. We have seen in Section 4 that there are a large number of markers and constructions that are somehow sensitive to oppositive contrast. Whether or not their opposition to other markers in the same language is stable or emergent, they do not cover exactly the same area of the oppositive contrast region of the similarity space, thus demonstrating by their cumulative cross-linguistic evidence that oppositive contrast is fuzzy. In languages with more than one relevant marker or construction, fuzziness can also be demonstrated within one language. This is illustrated here with Popti’, a language with several different phenomena sensitive to oppositive contrast.

Popti’ has basic VSO-order (Craig 1977: 8) with preverbal negation. A sentence-final first person clitic =an, glossed FIN.1, indicates that some subject, object or possessor in the sentence is first person. This marker also occurs at the end of the topic intonation unit (Aissen 1992: 61). Put differently, in some uses =an is a kind of selective cumulating with, and restricted to, first person. However, preverbal position of non-predicative phrases can also be triggered by the oppositive contrast marker wal (Craig 1977: 37), but this construction differs from topicalization in that pronouns can occur preverbally, as in (29), which is not allowed in topicalization (Craig 1977: 56).

(29) Popti’ (jac-x-bible 40026011)

Anma meba’ ay-ñetic’aco t-e-xol,
people poor EX-always PREFIX-2PL-among
wal-in=an xin mach sunilbal tiempo
but.OPP-ABS.1SG-FIN.1 then not all time
ay-in.ic’oj=an t-e-xol.
EX-ABS.1SG.here-FIN.1 PREFIX-2PL-among

‘for the poor you always have with you, but you do not always have me.’
Wal can be preceded by yaj/yaja ‘but’, but even bare yaj(a’) can attract initial NPs. In addition, the particle xin ‘then’ may follow an =an-clitic on an initial phrase (Day 1973: 97). While all these different phenomena are associated with initial non-predicative phrases, their distribution does not match exactly and cannot be explained by one single pre-established category such as a Top(ic)P(hrase)), testifying to the fuzziness of the oppositive contrast domain.

It may be assumed that initial oppositive phrases are the most important phenomenon type characterizing oppositive contrast cross-linguistically. Non-predicative initial phrases occur in an overwhelming majority of all VSO/VOS-languages considered in Section 4.2, and there is reason to believe that non-predicative initial phrases also occur in many other languages. The phenomenon just happens to be less salient in subject-initial languages where contrasted subjects are initial anyway. Oppositive contrast markers and selectives are more restricted in use: oppositive contrast markers have been identified in more than a seventh of the languages and selectives have so far been identified only in less than a tenth of the languages of the Bible corpus. Note that there are only very few exceptions to (27) “Oppositive contrast markers are immediately adjacent to oppositive phrases in prototypical oppositive contrast contexts”. This means that oppositive contrast marking will be induced by syntax to a very large extent. If many languages have some sort of oppositive contrast markers, as shown in Section 4.3, this is because various kinds of markers can assume this function if they happen to be associated with contrast-sentence-initial oppositive phrases. From this follows the assumption that many markers sensitive to oppositive contrast will not be dedicated to oppositive contrast in all of their functions, such as the particle Popti’ xin ‘then’ in (29).

A further illustrative example is the emphatic second position clitic že in Slavic languages which differs in its propensity to be associated with oppositive contrast from language to language. It is not formally ideal for signaling oppositive phrases, because it usually occurs after the first word. It is most strongly associated with oppositive contrast in Church Slavonic, where it is a calque from Koine Greek de(2), and in those Russian translations that are most strongly influenced by Church Slavonic. It is also common in Ukrainian (often reduced to ż, see example (14)). However, its borrowed version in Erzya Mordvin žo is rather strongly associated with oppositive contrast (high loading in PCA-analysis). Not incidentally, Erzya Mordvin žo is noun- and postposition-phrase-final (l’ija-tn’e-n’ marto žo [other-PL.DEF-GEN with but.OPP] ‘but with the others …’) 42008010), except if a relative clause follows. Put differently, the syntactic position of Erzya Mordvin žo affords oppositive contrast more easily than Slavic že.

We can conclude that although oppositive contrast can be conceived of as a semantic domain, it is syntactically co-conditioned. Oppositive contrast provides evidence that functional domains are not pre-established extralinguistic semantic
domains, but units reflecting non-arbitrary relationships between meaning and form.

This raises the question as to how non-arbitrary relationships between meaning and form should be explained, a question that can only be touched upon here. The most often adduced candidates are iconicity (Haiman 1985) and frequency, but these are not the only options. Gibson’s (1979) notion of affordance (see also Du Bois 2014: 367) is a further candidate. The notion of affordance implies that properties of objects can be immediately meaningful to observers without recourse to previous experience (such as a stone directly “affords” graspsability and throwability). For contrast constructions this means that an emphatic particle following a non-predicative initial phrase in a contrast sentence directly affords oppositive contrast, which can explain why oppositive contrast markers are often emergent (weakly conventionalized).

5.2 The consequences of fuzziness of the oppositive contrast domain

The results of this study show that the oppositive contrast domain is fuzzy with scalar contexts with unexpected oppositive phrases such as (30) reflecting the most prototypical examples. Contrast is a rhetorical discourse relation, which is why it is important to consider real corpus examples rather than isolated constructed examples such as (1). Not incidentally, all examples on the oppositive pole of the oppositive–counterexpectative dimension in Section 4.1 are from direct speech. The major discourse function of the most prototypical oppositive contrast examples in the NT is explanation where some sort of wise speaker corrects what is rhetorically framed as an addressee’s misconception. New common ground is intersubjectively established stepwise. Example (30) is part of Jesus’ explanation to Simon the Pharisee why a woman, who was known to be a sinner, acted in an impeccable way.

(30) English (42007046)

You did not anoint my head with olive oil, but she anointed my feet with perfumed oil.

Fuzziness of the domain is mainly due to degree of emphasis. Strength of emphasis presupposes degree of intensity on a scale. The most prototypical strong oppositive contrast contexts have multiple scales, such as the three scales in (30), shown in Table 8.
To summarize, gradualness of oppositive contrast between contrasted phrases is mainly rhetorical in character and highly context-dependent (cannot easily be addressed in isolated examples without context). It is strongly connected to degree of emphasis of oppositive phrases (see also Section 2.2).

### 5.3 Limitations and outlook

This study has many limitations. The functions that could be established on the oppositive–counterexpectative dimension (Table 5) are dependent on what kind of examples happen to occur in the NT. There are only few examples with strong oppositive contrast, such as (30), and they all happen to have a woman as the surprising oppositive referent (interestingly, oppositive contrast is also a relevant domain for incipient feminine gender markers, see Wälchli 2019: 65). Further research is needed to explore more details of the oppositive–counterexpectative distinction. However, if there are strong limitations on contexts available in the NT, this data source is more useful than reference grammars and dictionaries alone where examples are usually few and usually presented without discourse context.

This study focuses on a restricted region of the similarity space of contrast (for a larger picture, see Malchukov 2004). Concessive and restrictive contexts and contexts intermediate between contrast and coordination were not considered.

Most importantly, there is also a need for many language-specific studies that investigate oppositive contrast in original texts, taking into account that a large amount of language-internal variability may be expected. Very few markers sensitive to oppositive contrast have so far been subject to more detailed study, mainly from Indo-European languages, for instance Polish a (Andrason 2020) and Latin autem (Kroon 2019, ch. 10).
6 Conclusions

I will focus here subjectively on the major lessons I have learned myself, by carrying out this study. The most important point is the paramount methodological and theoretical relevance of fuzzy domains. Only in fuzzy domains can we see how diversity (absence of strict universals) and strong constraints do not exclude each other and how important it is to use statistical tools to identify major trends. In this respect, contrast is revealing, since the fuzzy oppositive contrast domain appears in sharp opposition to corrective contrast (Section 4.1), a more strictly delimited domain where the choice of method has less impact on what kind of results can be obtained. Even in fuzzy data it is most important to watch out for simplicity: for how empirically observable data is constrained.

Next, I was surprised by the strong and immediate interplay between form and function. Although oppositive contrast is a meaning, its manifestation is directly tied to certain word order patterns, clearly demonstrating that this functional domain cannot be abstracted from form (suggesting that functional domains are not extra-linguistic in general, but units of language). I further hope to have shown that word order in contrast constructions deserves more study.

Finally, many authors have emphasized the rhetorical and intersubjective character of contrast. I was surprised how right they are and that several crucial aspects of contrast constructions can only be detected in discourse contexts, which testifies to the importance of combining cross-linguistic and corpus-linguistic approaches.

List of abbreviations

= clitic
\~ reduplication
1 first person/initial position of contrast marker
2 second person/OP-final position of contrast marker
(2) second position clitic
3 third person
ABS absolutive
ACC accusative
ADD additive
ADESS adessive
AFF affirmative
AGT agentive
ALLAT allative
AP anchor phrase
AUX auxiliary
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEXP</td>
<td>counterexpectative</td>
</tr>
<tr>
<td>CONEG</td>
<td>connegative</td>
</tr>
<tr>
<td>COVERED</td>
<td>covered ground</td>
</tr>
<tr>
<td>DEF</td>
<td>definite</td>
</tr>
<tr>
<td>DEM</td>
<td>demonstrative</td>
</tr>
<tr>
<td>EMPH</td>
<td>emphatic</td>
</tr>
<tr>
<td>EX</td>
<td>existential</td>
</tr>
<tr>
<td>F</td>
<td>feminine</td>
</tr>
<tr>
<td>FIN</td>
<td>final particle</td>
</tr>
<tr>
<td>FUT</td>
<td>future</td>
</tr>
<tr>
<td>GEN</td>
<td>genitive</td>
</tr>
<tr>
<td>IMPERS</td>
<td>impersonal</td>
</tr>
<tr>
<td>INESS</td>
<td>inessive</td>
</tr>
<tr>
<td>INF</td>
<td>infinitive</td>
</tr>
<tr>
<td>INS</td>
<td>instrumental</td>
</tr>
<tr>
<td>IO</td>
<td>indirect object</td>
</tr>
<tr>
<td>IPPV</td>
<td>imperfective</td>
</tr>
<tr>
<td>IRR</td>
<td>irreals</td>
</tr>
<tr>
<td>LOC</td>
<td>locative</td>
</tr>
<tr>
<td>M</td>
<td>masculine</td>
</tr>
<tr>
<td>MOOD</td>
<td>mood</td>
</tr>
<tr>
<td>NEG</td>
<td>negation</td>
</tr>
<tr>
<td>NMLZ</td>
<td>nominalizer</td>
</tr>
<tr>
<td>NOM</td>
<td>nominative</td>
</tr>
<tr>
<td>N</td>
<td>neuter</td>
</tr>
<tr>
<td>OBL</td>
<td>oblique</td>
</tr>
<tr>
<td>OP</td>
<td>opposite phrase</td>
</tr>
<tr>
<td>OPP</td>
<td>opposite</td>
</tr>
<tr>
<td>PART</td>
<td>partitive</td>
</tr>
<tr>
<td>PERF</td>
<td>perfect</td>
</tr>
<tr>
<td>PFV</td>
<td>perfective</td>
</tr>
<tr>
<td>PL</td>
<td>plural</td>
</tr>
<tr>
<td>POSS</td>
<td>possession</td>
</tr>
<tr>
<td>PREFIX</td>
<td>prefix</td>
</tr>
<tr>
<td>PREP</td>
<td>preposition</td>
</tr>
<tr>
<td>PRS</td>
<td>present</td>
</tr>
<tr>
<td>PST</td>
<td>past</td>
</tr>
<tr>
<td>PTC</td>
<td>particle</td>
</tr>
<tr>
<td>PTCP</td>
<td>participle</td>
</tr>
<tr>
<td>REAL</td>
<td>realis</td>
</tr>
<tr>
<td>REFL</td>
<td>reflexive</td>
</tr>
<tr>
<td>REL</td>
<td>relative</td>
</tr>
<tr>
<td>REP</td>
<td>reportative</td>
</tr>
<tr>
<td>SBJ</td>
<td>subject</td>
</tr>
<tr>
<td>SG</td>
<td>singular</td>
</tr>
<tr>
<td>SUBR</td>
<td>subordinator</td>
</tr>
<tr>
<td>TMA</td>
<td>tense/mood/aspect</td>
</tr>
</tbody>
</table>
Acknowledgements: I am very grateful to Östen Dahl for having introduced me to Sæbo (2002). Östen also read several different versions of this work and contributed many useful comments. Peter Arkadiev helped me to get access to important literature. I would also like to thank Martin Haspelmath, Anna Sjöberg, four anonymous reviewers and the editors of LT, in particular Brigitte Pakendorf, for many valuable and intriguing comments and Angela Terrill at LT for her strong support. This work would not have been possible without thousands of people having been engaged in translating the New Testament to lesser known languages.

References


Rudolph, Elisabeth. 1996. Contrast. Adversative and concessive relations and their expressions in English, German, Spanish, Portuguese on sentence and text level. Berlin: de Gruyter.


Trotzke, Andreas. 2017. The grammar of emphasis. From information structure to the expressive dimension. Berlin: de Gruyter.

Tuuk, Hermanus Neubronner van der. 1864. Tobasche spraakkunst. Amsterdam: Muller.


**Supplementary Material:** This article contains supplementary material (https://doi.org/10.1515/lingty-2022-0019).